

# INF 511 Modern Regression Lecture Notes

© J. Barber

November 28, 2018



# Contents

|   |           |
|---|-----------|
| <b>1 Preliminaries &amp; Fundamental Concepts</b>               | <b>1</b>  |
| 1.1 Textbook Code/Data & Software Packages . . . . .            | 3         |
| 1.1.1 BFRM Web site . . . . .                                   | 3         |
| 1.1.2 R . . . . .   | 3         |
| 1.1.3 WinBUGS . . . . .   | 3         |
| 1.1.4 OpenBUGS . . . . .  | 4         |
| 1.1.5 JAGS . . . . .  | 4         |
| 1.1.6 Stan . . . . .  | 4         |
| 1.1.7 NIMBLE . . . . .  | 4         |
| 1.2 Introduction . . . . .                                      | 4         |
| 1.3 Random Mechanisms, Mixing-Up & Scope of Inference . . . . . | 9         |
| 1.4 Randomization Distribution . . . . .                        | 18        |
| 1.5 Sampling Distribution . . . . .                             | 33        |
| 1.6 Notes on Observational Studies . . . . .                    | 43        |
| 1.7 Unobserved Covariates and “Randomness” . . . . .            | 44        |
| 1.8 Subjective Probability . . . . .                            | 45        |
| 1.9 Summary . . . . .   | 45        |
| <b>2 Motivating Examples and a Course Outline</b>               | <b>47</b> |
| 2.1 Dental Growth . . . . .                                     | 49        |
| 2.2 Smoking and Forced Air Expiratory Volume (FEV1) . . . . .   | 50        |
| 2.3 Outcome After Head Injury . . . . .                         | 52        |
| 2.4 Contraception Drug . . . . .                                | 55        |
| 2.5 Seizure Data . . . . .                                      | 57        |
| 2.6 Lung Cancer & Radon . . . . .                               | 61        |
| 2.7 Aircraft Fastener Data . . . . .                            | 66        |
| 2.8 Cardiac Failure & Cadralazine Concentration . . . . .       | 71        |
| 2.9 Pharmacokinetics of Theophylline Data . . . . .             | 74        |
| 2.10 Course/Textbook Outline . . . . .                          | 77        |
| 2.10.1 Summary of This Course . . . . .                         | 80        |

|  |            |
|--|------------|
| <b>3 Basic Results in Probability and Statistics</b>                         | <b>81</b>  |
| 3.1 Summations & Products . . . . .  | 83         |
| 3.1.1 Summation Operator . . . . .   | 83         |
| 3.1.2 Double Summation Operator . . . . .                                    | 85         |
| 3.1.3 Product Operator . . . . .   | 85         |
| 3.2 Random Variables . . . . .   | 86         |
| 3.3 Characteristics of Random Variables . . . . .                            | 100        |
| 3.3.1 Expected (Mean) Value . . . . .  | 100        |
| 3.3.2 Variance Operator . . . . .  | 104        |
| 3.4 Random Vectors . . . . .   | 106        |
| 3.4.1 Covariance Operator & Its Properties . . . . .                         | 106        |
| 3.4.2 Independence . . . . .   | 109        |
| 3.5 Central Limit Theorem . . . . .  | 110        |
| 3.6 Linear Functions of an RV . . . . .                                      | 111        |
| 3.7 Linear Combinations of RVs . . . . .                                     | 112        |
| <b>4 Matrices &amp; Vectors</b>  | <b>115</b> |
| 4.1 Notation, Dimension, Rows, Columns, Elements . . . . .                   | 117        |
| 4.2 Matrix Arithmetic . . . . .  | 121        |
| 4.2.1 Addition/Subtraction . . . . .   | 122        |
| 4.2.2 Multiply a Matrix by a Scalar . . . . .                                | 124        |
| 4.2.3 Matrix Multiplication . . . . .  | 124        |
| 4.2.4 Matrix Transpose . . . . .   | 126        |
| 4.2.5 Special Matrices . . . . .   | 129        |
| 4.2.6 Linear Dependence & Rank . . . . .                                     | 132        |
| 4.3 Combining Things: Random Vectors and Matrices . . . . .                  | 141        |
| 4.3.1 Expectation of a Random Vector/Matrix . . . . .                        | 142        |
| 4.3.2 Variance(-Covariance) Matrix . . . . .                                 | 142        |
| 4.3.3 Linearity of Expectation Operator (just as in scalar case) . . . . .   | 143        |
| 4.3.4 Variance(-Covariance) of $\mathbf{a} + \mathbf{B}\mathbf{Y}$ . . . . . | 143        |
| 4.3.5 Distribution of Linear Function of Normal RV . . . . .                 | 143        |
| <b>5 Linear Models I: Introduction</b>                                       | <b>149</b> |
| 5.1 Motivating Data Set . . . . .  | 151        |
| 5.2 Distributions: Joint, Marginal & Conditional . . . . .                   | 153        |
| 5.3 Model Specification . . . . .  | 160        |
| 5.3.1 It's a Conditional Mean Model Specification . . . . .                  | 161        |
| 5.3.2 Covariates Observed Without Error . . . . .                            | 163        |
| 5.4 A Justification of Linear Modeling . . . . .                             | 163        |
| 5.5 Parameter Interpretation . . . . .                                       | 165        |
| 5.5.1 Conditional Mean Model vs. Marginal Mean model . . . . .               | 165        |
| 5.5.2 Extrapolation, Meaningful Parameters & Reparameterization . . . . .    | 166        |
| 5.5.3 Typical "Additive Change" Parameter Interpretation . . . . .           | 168        |

|          |   |            |
|----------|---|------------|
| 5.5.4    | Data Transformations . . . . .  | 173        |
| <b>6</b> | <b>Linear Models II: Frequentist Approach</b>                                       | <b>175</b> |
| 6.1      | General Linear Model . . . . .  | 177        |
| 6.2      | (Ordinary) Least Squares . . . . .  | 179        |
| 6.3      | Gauss–Markov Theorem . . . . .  | 182        |
| 6.3.1    | Remarks on the Gauss Markov Theorem . . . . .                                       | 183        |
| 6.4      | Normal Maximum Likelihood . . . . .   | 184        |
| 6.5      | Fitted Values, Residuals, Hat Matrix and MSE . . . . .                              | 187        |
| 6.6      | Distributions Following from the Normal Linear Model . . . . .                      | 189        |
| 6.6.1    | $z$ and $\chi^2$ Distribution Results . . . . .                                     | 189        |
| 6.6.2    | Standard Error of an Estimator . . . . .  | 194        |
| 6.6.3    | $t$ and $F$ Distribution Results . . . . .  | 195        |
| 6.6.4    | Estimated Standard Error of an Estimator . . . . .                                  | 197        |
| 6.6.5    | Summary of Distributional Results . . . . .   | 198        |
| 6.7      | Tests and Intervals using $t$ and $F$ Distribution Results . . . . .                | 199        |
| 6.7.1    | Two Basic Questions . . . . .   | 199        |
| 6.7.2    | General Linear Hypothesis . . . . .   | 200        |
| 6.7.3    | $\mathbf{C}\boldsymbol{\beta}$ Approach with $F$ . . . . .                          | 200        |
| 6.7.4    | Full vs. Reduced Model or Extra Sum-of-Squares Approach . . . . .                   | 200        |
| 6.7.5    | Special Case: Omitting Some Variables or Setting Some $\beta_j$ to 0 . . . . .      | 202        |
| 6.7.6    | $t$ tests for Scalar $\mathbf{C}\boldsymbol{\beta}$ . . . . .                       | 203        |
| 6.7.7    | $t$ tests for Omitting a Single $x_j$ or Setting $\beta_j = 0$ . . . . .            | 203        |
| 6.7.8    | $t$ -based Confidence Intervals for Scalar $\mathbf{C}\boldsymbol{\beta}$ . . . . . | 204        |
| 6.7.9    | $t$ -based Prediction Intervals for $Y   \mathbf{x}$ . . . . .                      | 205        |
| 6.8      | Example Data Analysis Using $t$ and $F$ Results . . . . .                           | 206        |
| 6.9      | Summary and Final Remarks . . . . .   | 215        |
| <b>7</b> | <b>Linear Models III: Example Frequentist Data Analysis</b>                         | <b>217</b> |
| 7.1      | Introduction . . . . .  | 219        |
| 7.1.1    | Overview . . . . .  | 219        |
| 7.1.2    | Preview . . . . .   | 221        |
| 7.2      | Data Set . . . . .  | 222        |
| 7.3      | Graphical Exploratory Data Analysis (EDA) . . . . .                                 | 225        |
| 7.4      | Initial Linear Regression Model . . . . .   | 233        |
| 7.5      | Omitted & Added Variable Plots . . . . .  | 234        |
| 7.6      | Remodeling . . . . .  | 237        |
| 7.7      | Observed vs. Fitted Plot . . . . .  | 238        |
| 7.8      | Residual Plots . . . . .  | 239        |
| 7.9      | Overall F Test: Special Case of $\mathbf{C}\boldsymbol{\beta}$ . . . . .            | 242        |
| 7.9.1    | <code>anova</code> Function for F v R Approach . . . . .                            | 244        |
| 7.9.2    | The <code>glh.test</code> R Function . . . . .                                      | 247        |
| 7.10     | $R^2$ & Adjusted $R^2$ . . . . .  | 248        |

|          |  |            |
|----------|--|------------|
| 7.11     | Using an Interaction Term . . . . .  | 250        |
| 7.12     | $t$ -based inference for $\beta_j$ . . . . .   | 257        |
| 7.12.1   | Default lm Printout and summary . . . . .  | 257        |
| 7.12.2   | By Hand Test and Intervals . . . . .   | 258        |
| 7.12.3   | The confint Function . . . . .   | 259        |
| 7.12.4   | The estimable Function . . . . .   | 259        |
| 7.13     | Qualitative Covariates . . . . .   | 260        |
| 7.13.1   | Cell Reference Coding . . . . .  | 260        |
| 7.14     | Intervals for $E(Y   \mathbf{x})$ & $Y   \mathbf{x}$ with predict . . . . .              | 266        |
| 7.14.1   | Another Warning About Extrapolation . . . . .  | 268        |
| 7.15     | Summary . . . . .  | 269        |
| <b>8</b> | <b>Bayesian Linear Model</b>   | <b>271</b> |
| 8.1      | Introduction . . . . .   | 275        |
| 8.2      | Distributions: Data, Prior & Posterior . . . . .   | 275        |
| 8.2.1    | Data Distribution . . . . .  | 276        |
| 8.2.2    | Bayes Theorem: Data, Prior, Joint, Marginal & Posterior . . . . .                        | 276        |
| 8.3      | Summary So Far . . . . .   | 283        |
| 8.4      | Linear Model . . . . .   | 284        |
| 8.5      | Conjugate Prior . . . . .  | 285        |
| 8.5.1    | Posterior . . . . .  | 286        |
| 8.5.2    | Marginal Posterior for $\boldsymbol{\beta}$ is a $t$ . . . . .                           | 288        |
| 8.5.3    | Posterior Predictive is a $t$ . . . . .  | 290        |
| 8.5.4    | Remarks . . . . .  | 291        |
| 8.6      | A Common Improper Prior . . . . .  | 292        |
| 8.6.1    | Posterior . . . . .  | 293        |
| 8.6.2    | Marginal Posterior for $\boldsymbol{\beta}$ is a Familiar $t$ . . . . .                  | 293        |
| 8.6.3    | Posterior Predictive is a Familiar $t$ . . . . .   | 294        |
| 8.7      | STAT 101 Redux a la Bayes . . . . .  | 295        |
| 8.7.1    | $t$ -based Intervals for $\beta_j$ . . . . .   | 296        |
| 8.7.2    | $t$ -based Test for $\beta_j$ . . . . .  | 297        |
| 8.7.3    | $t$ -based Intervals for $E(Y   \mathbf{x}) = \mathbf{x}^t \boldsymbol{\beta}$ . . . . . | 299        |
| 8.7.4    | $t$ -based Prediction Intervals for $Y   \mathbf{x}$ . . . . .                           | 300        |
| 8.8      | Example . . . . .  | 301        |
| 8.8.1    | Frequentist R Summary . . . . .  | 301        |
| 8.8.2    | Bayesian Summary . . . . .   | 302        |
| 8.9      | A Common Independence Prior . . . . .  | 307        |
| 8.9.1    | Full Conditional Posterior Distributions . . . . .                                       | 308        |
| 8.10     | 2-Stage Gibbs Sampling . . . . .   | 309        |
| 8.11     | Example . . . . .  | 309        |
| 8.11.1   | Eliciting a Prior . . . . .  | 310        |
| 8.12     | Hamiltonian Monte Carlo in Stan . . . . .  | 314        |
| 8.12.1   | Functions Block . . . . .  | 314        |

| Lecture 0  | Page v     |
|--|------------|
| 8.12.2 Data Block . . . . .  | 314        |
| 8.12.3 Transformed Data Block . . . . .                                      | 315        |
| 8.12.4 Parameters Block . . . . .  | 315        |
| 8.12.5 Transformed Parameters Block . . . . .                                | 315        |
| 8.12.6 Model Block . . . . .   | 316        |
| 8.12.7 Generated Quantities Block . . . . .                                  | 316        |
| 8.12.8 Altogether for Stan . . . . .   | 316        |
| 8.12.9 Translate Stan to C++ with <code>stanc</code> . . . . .               | 318        |
| 8.12.10 Make an Executable Stan Model with <code>stan_model</code> . . . . . | 318        |
| 8.12.11 Data List for Stan . . . . .   | 319        |
| 8.12.12 List of Initial Value Lists for Stan . . . . .                       | 319        |
| 8.12.13 Executing a Stan Model with <code>sampling</code> . . . . .          | 320        |
| 8.12.14 Posterior Summaries with <code>coda</code> . . . . .                 | 320        |
| 8.13 Example Summary . . . . .   | 325        |
| 8.14 Other Priors . . . . .  | 327        |
| <b>9 One-Way ANOVA</b>   | <b>329</b> |
| 9.1 Initial Concepts and Notation . . . . .                                  | 331        |
| 9.2 Cell Means (Regression) Model: $E(Y_{ij}) = \mu_i$ . . . . .             | 333        |
| 9.2.1 Example . . . . .  | 337        |
| 9.3 Cell Means Model: Further Inference About Means . . . . .                | 344        |
| 9.3.1 Example . . . . .  | 345        |
| 9.4 Factor Effects Parameterization . . . . .                                | 347        |
| 9.4.1 Defining a Factor Effects Parameterization Using Treatment Means .     | 348        |
| 9.5 Factor Effects Parameterization: Before Constraints . . . . .            | 349        |
| 9.6 Imposing Constraints . . . . .   | 350        |
| 9.7 Sum-to-Zero Constraint/Coding . . . . .                                  | 351        |
| 9.8 Reference Treatment Constraint/Coding . . . . .                          | 360        |
| 9.9 Further Inference About Treatment Means . . . . .                        | 368        |
| 9.9.1 Example . . . . .  | 369        |
| 9.10 Summary of One-Way ANOVA . . . . .                                      | 373        |
| 9.10.1 Regression Approach to ANOVA . . . . .                                | 376        |
| 9.10.2 Model, Parametrization, Reparameterization . . . . .                  | 376        |
| <b>10 Multi-Way ANOVA</b>  | <b>379</b> |
| 10.1 Initial Concepts and Notation . . . . .                                 | 381        |
| 10.2 ANOVA Model Components: Means and Effects . . . . .                     | 383        |
| 10.3 Example . . . . .   | 390        |
| 10.4 Cell Means Model of $E(Y_{ijk})$ . . . . .                              | 395        |
| 10.5 Factor Effects Parameterization: Before Constraints . . . . .           | 396        |
| 10.6 Factor Effects: Sum-to-Zero Constraints/Coding . . . . .                | 400        |
| 10.6.1 E.g.: Factor Effects S2Zero Initial Analysis . . . . .                | 404        |
| 10.6.2 E.g.: Effects S2Zero ANOVA For Common $\mathbf{C}\beta$ . . . . .     | 419        |

|  |     |
|--|-----|
| 10.6.3 E.g.: Factor Effects S2Zero F v R & $\mathbf{C}\boldsymbol{\beta}$ Approach . . . . . | 423 |
| 10.6.4 E.g.: Effects 2Zero Summary . . . . .   | 425 |
| 10.7 Factor Effects: Treatment Constraints/Coding . . . . .                                  | 425 |
| 10.7.1 E.g.: Factor Effects Trmt Initial Analysis . . . . .                                  | 430 |
| 10.7.2 E.g.: Effects Trmt ANOVA For Common $\mathbf{C}\boldsymbol{\beta}$ . . . . .          | 435 |
| 10.7.3 E.g.: Factor Effects Trmt F v R & $\mathbf{C}\boldsymbol{\beta}$ Approach . . . . .   | 436 |
| 10.7.4 E.g.: Effects Trmt Summary . . . . .  | 438 |
| 10.8 SS Type, Balance & the Marginality Principle . . . . .                                  | 438 |
| 10.8.1 Sequential SS ANOVA . . . . .   | 440 |
| 10.8.2 Partial SS ANOVA . . . . .  | 443 |
| 10.8.3 Marginality Principle . . . . .   | 444 |
| 10.8.4 Balance . . . . .   | 448 |
| 10.9 Additive Model: Tests for Overall Main Effects . . . . .                                | 449 |
| 10.9.1 F v R Approach . . . . .  | 449 |
| 10.9.2 $\mathbf{C}\boldsymbol{\beta}$ Approach . . . . .                                     | 452 |
| 10.10 Additive Model: More Detailed Inference of Main Effects . . . . .                      | 453 |

# List of Tables



# List of Figures

|     |  |     |
|-----|--|-----|
| 1.1 | Scope of inference . . . . .   | 10  |
| 3.1 | Continuous cdf. . . . .  | 88  |
| 3.2 | Relationship between pdf and cdf. . . . .                                | 90  |
| 3.3 | Z-table (source: [KNNL05, Table B.1]). . . . .                           | 92  |
| 3.4 | Relationship between pmf and cdf. . . . .                                | 95  |
| 3.5 | Table of cumulative binomial probabilities (Source: forgotten!). . . . . | 97  |
| 3.6 | Discrete uniform pmf supported on 1, 2, 3. . . . .                       | 102 |
| 3.7 | pdf is balanced on mean ( $\mu$ ). . . . .                               | 103 |
| 3.8 | Sketch to illustrate covariance. . . . .                                 | 107 |
| 3.9 | Sketch of marginal and joint distributions. . . . .                      | 110 |
| 5.1 | Bivariate regression model . . . . .                                     | 162 |
| 6.1 | OLS Geometry . . . . .   | 181 |



# Lecture 1

## Preliminaries & Fundamental Concepts

### Contents

---

|            |  |           |
|------------|--|-----------|
| <b>1.1</b> | <b>Textbook Code/Data &amp; Software Packages . . . . .</b>            | <b>3</b>  |
| 1.1.1      | BFRM Web site . . . . .  | 3         |
| 1.1.2      | R . . . . .  | 3         |
| 1.1.3      | WinBUGS . . . . .  | 3         |
| 1.1.4      | OpenBUGS . . . . .   | 4         |
| 1.1.5      | JAGS . . . . .   | 4         |
| 1.1.6      | Stan . . . . .   | 4         |
| 1.1.7      | NIMBLE . . . . .   | 4         |
| <b>1.2</b> | <b>Introduction . . . . .</b>  | <b>4</b>  |
| <b>1.3</b> | <b>Random Mechanisms, Mixing-Up &amp; Scope of Inference . . . . .</b> | <b>9</b>  |
| <b>1.4</b> | <b>Randomization Distribution . . . . .</b>                            | <b>18</b> |
| <b>1.5</b> | <b>Sampling Distribution . . . . .</b>                                 | <b>33</b> |
| <b>1.6</b> | <b>Notes on Observational Studies . . . . .</b>                        | <b>43</b> |
| <b>1.7</b> | <b>Unobserved Covariates and “Randomness” . . . . .</b>                | <b>44</b> |
| <b>1.8</b> | <b>Subjective Probability . . . . .</b>                                | <b>45</b> |
| <b>1.9</b> | <b>Summary . . . . .</b>   | <b>45</b> |

---

*Additional Reading:*

- [Wak13, Chap. 1 except §1.3 for now]

---

 $\mathcal{R}$ 

### ***Main Objectives:***

- Inference.
- Statistical inference formally accounts for uncertainty via random mechanisms.
- Scope of inference is founded in random mechanisms.
- Randomized experiments justify causal inference.
- Observational studies do not justify causal inference (without further assumptions).
- Randomization distributions from randomized experiments.
- Random sampling justifies inference to populations.
- Sampling distributions from random samples from populations.
- Central Limit Theorem (CLT)
- Notion of hypothetical replications is fundamental to randomized experiments, random sampling, and to frequentist statistics interpretation of probability in general.
- Unmeasured variables are often modeled as normal errors via the CLT and can alter interpretation of effects.
- Subjective interpretation of probability.
- Introduction to R and other computing/online resources.

---

 $\mathcal{O}$

- RECALL, FROM OUR SYLLABUS, THAT THESE NOTES ARE NOT COMPLETE (OR EVEN CORRECT!) BUT WILL BE SUPPLEMENTED (AND CORRECTED!) IN CLASS. IN OTHER WORDS, COME TO CLASS AND BE PREPARED TO TAKE ADDITIONAL NOTES.
- *SIGN THE ATTENDANCE SHEET!*

## 1.1 Textbook Code/Data & Software Packages

Your text's author focuses nearly exclusively on R and its packages and on WinBUGS, the latter of which is no longer officially developed, so I list a few other Bayesian probabilistic programming languages that you may want to adapt the text's WinBUGS code to. I will tend to use Stan, perhaps exclusively.

All of these softwares are free. You are expected to make arrangements yourselves to use these, whether you find these installed on NAU IT lab computers or you install them on your own personal computer. Having said that, get R and Stan (via one of Stan's several interfaces, e.g., RStan).

### 1.1.1 BFRM Web site

<http://faculty.washington.edu/jonno/regression-methods.html>

### 1.1.2 R

<https://www.r-project.org>

### 1.1.3 WinBUGS

<https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/>

### 1.1.4 OpenBUGS

<http://www.openbugs.net/w/FrontPage>

### 1.1.5 JAGS

<http://mcmc-jags.sourceforge.net>

### 1.1.6 Stan

<http://mc-stan.org>

### 1.1.7 NIMBLE

(Numerical Inference for statistical Models for Bayesian and Likelihood Estimation)

<https://r-nimble.org>

## 1.2 Introduction

- This remainder of this first chapter of our lecture notes is based mostly on [RS13, Chapter 1], which contains relatively simple but fundamentally important concepts from the field of statistics.
- These concepts serve to justify generalizations from data to a **broader context, beyond the data at hand**.
- In other words, these concepts are fundamental to making **inference** and hence to science.
- In **later** chapters and throughout the course, we will also discuss how **statistical methods and their assumptions** justify (or not) conclusions from our data.
- Thus, knowing about data and methods help to **justify conclusions** drawn from data and methods.

- You will find related discussions in [Wak13, §1.2 & 1.4] (our required text) and [KNNL05, §9.1 & Chap. 15].
- We use this Introduction to introduce a few definitions, to be discussed in class.

### Definition 1.1 (Experiment).

- *A process of obtaining an observation / measurement / outcome / response from an object.*
- *Note that this definition is very broad relative to its more typical meaning in, say, “randomized experiments,” about which we will have a bit more to say, shortly.*

### Definition 1.2 (Unit).

- *An object from which an observation / measurement / outcome / response is obtained in an experiment.*
- *Context will dictate use of other, relatively common and synonymous terms such as **observational unit**, **experimental unit**, **sampling unit**, **subject**, **participant** (and then some!).*

### Definition 1.3 (Variable).

- *An observation / measurement / outcome / response obtained from a unit in an experiment.*

- A variable may be (i) **categorical** (non-numerical) or (ii) **numerical**.
- Categorical variables are often numerically coded, although the meaning of the numbers used to code categorical variables is usually not the same as the numbers associated with numerical variables.
- E.g., hair color may be treated as a categorical variable with different values black, blonde, brown, red, and other, which may be coded numerically as, e.g., 0,1,2,3,4. But, in this case, the numbers are not meant to imply any order or any meaning to differences or ratios; what would  $2-1=1$  or  $2/1=2$  mean here? Nothing. Is brown one more than (or twice) blonde?

---

**Definition 1.4** (Random Variable).

- A variable that cannot be predicted with certainty before it is observed.
- In other words, a variable whose value is uncertain before it is observed.
- There are more technical definitions of a random variable, but I do not see how such definitions serve us.

---

**Definition 1.5** (Data).

- A collection of observed random variables.
- A **sample**.

**Definition 1.6** (Inference).

- *An inference is a conclusion that patterns in data are present in some broader context.*
- *Or, inference is the process of drawing such conclusions.*

**Definition 1.7** (Statistical Inference).

- *A statistical inference is an inference that is justified by a probability model that links data to a broader context.*
- *Such probability models arise from the use of random mechanisms (or via theory and assumptions), to be discussed, shortly.*

**Definition 1.8** (Probability).

- *We do not give a formal definition of probability. Instead, we will consider a more intuitive development, mostly through examples, using a few properties of probability along the way (e.g.,  $\Pr(X < -2 \text{ OR } X > 2) = \Pr(X < -2) + \Pr(X > 2)$  (disjoint sets), as you likely computed in a previous class using a normal or t probability distribution).*
- *In any case, I am compelled to state one important defining property of probabilities: they are numbers **between zero and one** (inclusive)!*

**Definition 1.9** (Treatment).

- A treatment is the value of a variable, often categorical, that is assigned to a unit. E.g., drug variable values of “drug” and “no drug”.
- Often, treatments are defined by the combination of the values of two or more variables (often called **factors**). E.g., treatment of humidity = 50% and temperature = 25 degrees celcius.
- Typically, we are interested in the **relationship** of an outcome variable with treatments (or with factors); we often hear “what is the treatment (factor) effect (on the response/outcome variable)?”
- Sometimes, “treatment” (or factor) may be used even when such conditions are not assigned to units but are instead **inherent to the units**. E.g., gender, ethnicity, soil type, location, etc.
- We will have more to say about treatments and factors when we get to analysis of variance (ANOVA), later.

**Definition 1.10** (Covariate).

- A variable that is thought to be associated with an outcome (response) variable. (a.k.a, regressor, predictor, input, independent, ...)
- Often quantitative, as in regression, but less often may refer to a qualitative variable (factor), as in ANOVA.

**Definition 1.11** (Effect).

- The effect of a variable on another variable.

- *Very often (not always) specified explicitly as a parameter in a statistical model (because we are often interested explicitly in the effect of variables)*
- *You often hear ‘treatment effect’ (for a factor level) or ‘regression effect’ (for a quantitative covariate).*
- *Care must be taken when interpreting effects as causal, i.e., as if a change in one variable causes a change in another variable.*

### 1.3 Random Mechanisms, Mixing-Up & Scope of Inference

- How data may or may not allow us to justifiably generalize beyond the data, i.e., to infer, depends on how data are produced.
- We will refer to [RS13, Display 1.5, page 9] (Figure 1.1) throughout much of this chapter.

- **Two fundamental random mechanisms** (or variations thereof) underly the construction of **probability models** that serve statistical inference.
  1. The **randomization of units** to (treatment) groups (or vice-versa) defined by common conditions (i.e., treatment levels) set by an experimenter. This randomization is the defining concept of **randomized experiments** and **causal inference**. This situation is depicted by the left column of [RS13, Display 1.5, page 9], which is reproduced in Figure 1.1.

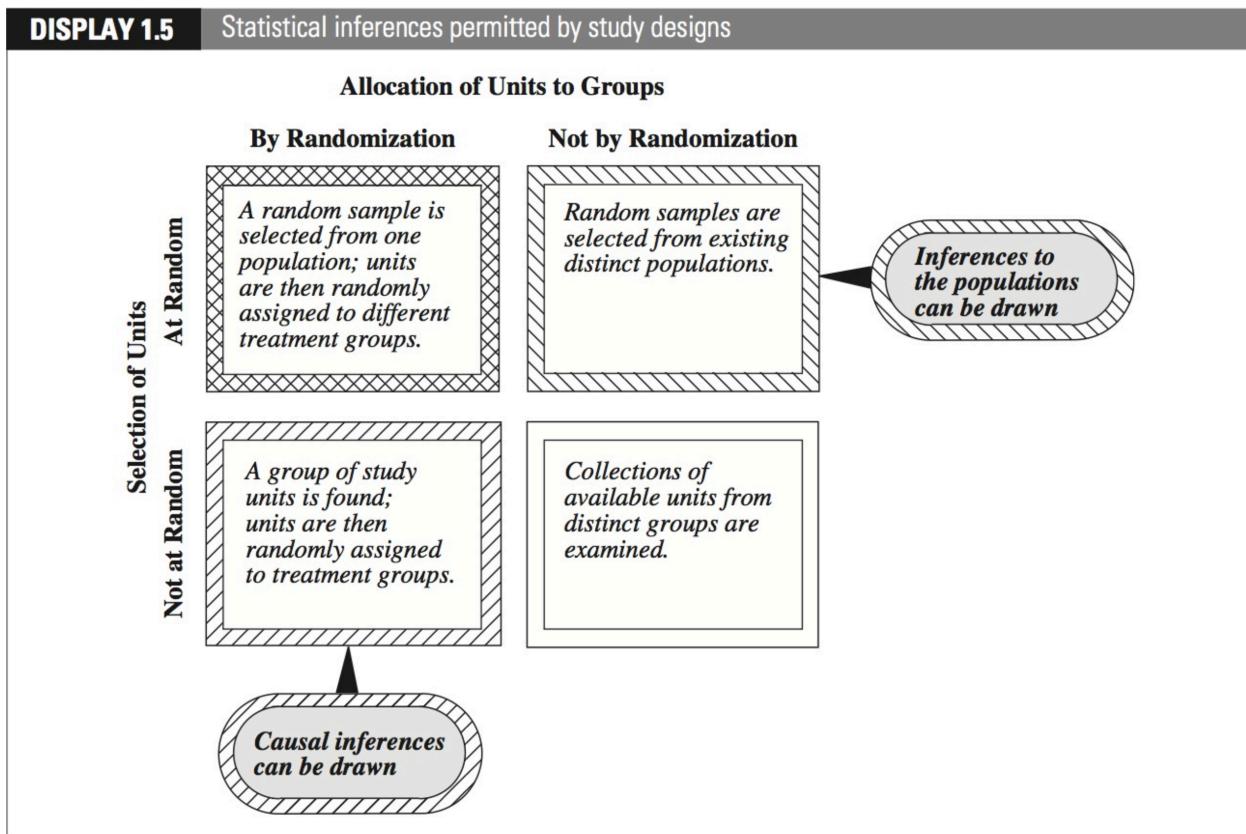


Figure 1.1: Scope of inference (Source: [RS13]).

2. The random selection, i.e., **random sampling**, of a subset of units from some larger collection of units about which we wish to infer. This random selection justifies inference from a **sample** of units to a **population** of units. This situation is depicted by the top row of [RS13, Display 1.5, page 9] (1.1).
- Without such random mechanisms, scope of inference is limited. This is depicted by the lower right box of [RS13, Display 1.5, page 9] (1.1).

---

**Definition 1.12** (Randomized Experiment).

- The first mechanism, above, including many variants thereof, is the defining characteristic of **randomized experiments** and is the most widely accepted mechanism to justify an inference of a **cause and effect** relationship between variables measured on units.*
- You may think of the “inference,” in this case, being that a **causal** relationship somehow serves as an **underlying fundamental explanation** for what we would otherwise call an **association** or **correlation** among variables. In this sense, randomization allows the **scope of inference** to go beyond the data at hand to infer an underlying mechanistic relationship between variables (e.g., treatment conditions and a response).*
- You may have heard, “correlation is not causation” (randomized experiments excepted).*
- Of course, we might casually (not causally!) attribute an association between variables to a cause-effect relationship, but it is the randomization of units to treatments that puts such an inference on firm ground statistically because we are then **able to formally characterize the chance of making incorrect conclusions of causal relationships**; a bit more on this, below.*

- Again, we are in the left column of [RS13, Display 1.5, page 9] (Figure 1.1).
- We will not cover randomized experiments (“experimental design”) much in this class.

**Remark 1.1** (Field of Causal Inference).

- As we will discuss, shortly, randomization is fundamental to concluding a cause-effect (causal) relationship between variables.
- However, a randomized experiment is not always convenient or possible, and we very often find ourselves with observational data.
- If we make **assumptions about the nature of causality**, then we may make causal inferences from observational data.
- Such causal inferences are the subject of the developing **field of Causal Inference** (e.g., **causal diagrams**, **path analysis**, **propensity scores**, **potential outcomes**, **interventions**, **counterfactuals**) ([Pea09], [PM18]).
- We will largely ignore the field of causal inference in this course, but this should not be taken to mean that the field of causal inference has nothing to offer!

**Definition 1.13.** *Observational Study*

- Studies that do not consist of randomized experiments, i.e., do not randomly assign units to treatment groups. Importantly, causal inferences are essentially not justified from observational study data; our scope of inference is relatively limited.

- See, however, our previous remark (1.1) about the Field of Causal Inference.
- Again, this is depicted in the right column of [RS13, Display 1.5, page 9] (Figure 1.1).

- For the second mechanism, **random sampling**, we introduce a few more definitions and remarks.
- We are now in the top row of [RS13, Display 1.5, page 9] (Figure 1.1).

#### Definition 1.14 (Population).

- The complete set of units of interest. We may denote the total number of units in a population (size) as  $N$  (though we do not consider population size much in this class).
- The population may not be well defined.
- In nearly all practical situations, this set of units is almost always beyond us, in some sense, but, instead, we focus on a subset of the population (sample).
- We will introduce the notion of a **superpopulation**, below ([Wak13, §1.2]), as an approximating model of the population.

**Remark 1.2** (Population of Units or Population of Values?). Note that some people (e.g., statisticians) may prefer to discuss sample and population in terms of the values of the (random) variables (outcomes) that

*may be observed from units—a sample (data)/population of values rather than sample/population of units. We may say something about this below, when we talk of “superpopulation,” but we will otherwise tend not to make a distinction between a population of units or a population of their values.*

**Definition 1.15** (Sample).

- *A subset of units from a population (if well defined). Data set. We often denote the total number of units in a sample (**sample size**) as  $n$ .*

**Definition 1.16** (Simple Random Sample).

- *A simple random sample of size  $n$  from a population is a subset of the population consisting of  $n$  members selected in such a way that every subset of size  $n$  is afforded the same chance of being selected.*

**Remark 1.3** (We Will Not Cover Sampling Methods).

- *In this course, we do not treat the many different ways to obtain samples from populations, but largely leave this important and rich area to other courses.*
- *However, random sampling from a population is fundamental to statistical inference of results, obtained from a sample, to a population, analogous to how randomization is fundamental to inferring causal relationships.*

- What is the underlying rationale for how randomization of units to treatment groups justifies causal inference?
- Similarly, what is the underlying rationale for how random sampling justifies conclusions about the population from a subset of observations?
- Before we attempt answers, we need another definition.

**Definition 1.17** (Confounding Variable).

- A *confounding variable* is a variable that is related to both (treatment) group/level membership (or to a covariate) (i.e., “ $x$ ” variable) and to an outcome variable (i.e., response or “ $y$ ” variable).
- Thus, an apparent association between group membership (or covariate) and an outcome may not be due to a direct (causal) relationship between group membership (or covariate) and an outcome, but to their shared relationship with other, confounding variables, which we often do not or cannot observe.
- In other words, confounding variables can make causal inference problematic.
- **Lurking variable** or **unobserved/unmeasured covariate** or **confounder** are frequently used synonyms.
- For example, though people living in households with more televisions may tend to be healthier than those living with fewer televisions, we do not say that televisions cause people to be more healthy (or that better health causes more televisions); in this case, factors such as diet, exercise and health care may be confounding factors associated with both health outcomes and material wealth, like TVs.

**Remark 1.4** (Mixing-Up Confounding Variables).

- Now back to our questions.
- The **randomization of units** to treatment groups, in a randomized experiment, “mixes-up” units among groups, hence mixes up the (unobserved) values of units’ confounding variables so that these confounding variables **tend** not to exhibit any systematic effect (**bias**) on units’ outcomes among treatment groups.
- In other words, for example, the proportion of “high responding” units, “low responding” units, “typical units” or “extreme” units **tends** to be the same, after group assignment, as is was before group assignment.
- Thus, if we do see a systematic effect among treatment groups and outcomes, we **tend** to think that the association is due to a direct (cause-effect) relationship between group membership (treatments) and the outcome, and not to confounding variables.
- For example, to illustrate (somewhat unrealistically), suppose that we want to study the relationship between blood cholesterol in human subjects and subjects’ cholesterol drug status: takes drug or does not take drug. It could be that subjects’ outcomes (cholesterol levels) are related to e.g., gender, but not to their drug status. (Granted, gender is usually observable!)
- Analogously, **random sampling** tends to result in samples whose proportions of confounding variable values are the same as their proportions in the population.
- In other words, the proportion of “high responding” units, “low responding” units, “typical units” or “extreme” units in the sample **tends** to be the same as in the population so that that the sample **tends** not to exhibit any systematic effects (**bias**), due to confounding variables, on sample units’ outcomes compared to the population.

- Thus, of course, we tend to think that results obtained from a sample reflect characteristics of the population. In this sense, we are statistically justified in inferring results from the sample to the population.

**Remark 1.5** (Hypothetical Replications).

- But you might object: by chance, the particular random assignment of units to treatments that we have to work with may not exhibit such mixing-up of confounders (and we won't know it because confounding variables are not typically observed).
- This is a good point to which the standard (frequentist, at least) response is to view our particular randomization as only one realization of a (usually) large number of possible **hypothetical replications** that could have occurred. Collectively, these (hypothetical) replications lead to a null/reference probability distribution (**randomization distribution**, below), which allows us to characterize the probability of wrongfully concluding a treatment effect when it doesn't exist (Type I error).
- Analogously, for a random sample: by chance, the particular random sample of units that we have to work with may not exhibit such such mixing-up of confounders (and we won't know it because confounding variables are not typically observed).
- Again, this is a good point to which the standard (frequentist, at least) response is to view our particular random sample as only one realization of a (usually) large number of possible **hypothetical replications** that could have occurred. Collectively, these (hypothetical) replications lead to a null/reference probability distribution (**sampling distribution**, below), which allows us to characterize the probability of making wrongful conclusions about the population(s) (Type I error).

**Remark 1.6** (Two Take-Home Messages for Scope of Inference).

1. *Statistical inferences of cause-and-effect relationships can be drawn from randomized experiments, but not from observational studies, at least not without further assumptions (see remark (1.1) about the Field of Causal Inference, above).*
2. *Inferences to populations can be drawn from random sampling studies, but not otherwise.*

## 1.4 Randomization Distribution

- If it's not already obvious, in a randomized experiment, we typically want to (somehow) compare responses among treatment levels to see if there is a "difference" or a "treatment effect."
- Here, we discuss how the **hypothetical replications** of a randomized experiment lead us to a **reference distribution (null distribution)** called the **randomization distribution**, which we will use to help us discover (test) if there is an effect.
- (Similarly, in a subsequent section, we will discuss a reference (null) distribution (**sampling distribution**), that arises from the hypothetical replication of random sampling.) But, first, we discuss the randomization distribution.
- For each of the **hypothetical replications** resulting from the random assignment of units to treatments in a **randomized experiment** (including the one assignment that you actually have), we can (hypothetically) observe values from our units.

**Example 1.1** (Consider First a Single Unit).

- Consider (unrealistically) a single ( $n=1$ ) unit to be assigned to one of  $T = 2$  treatments.
- Denote

$$Y_1$$

as the outcome/response of this unit 1 under [treatment 1], and

$$Y_1 + \delta$$

as the outcome of unit 1 under [treatment 2].

- Note that we have assumed that the effect of treatment level 2, relative to level 1, is **additive**, and we call

$$\delta$$

an **additive effect**, which is by far the most common modeling assumption for how treatments affect responses; we will use it almost without thinking for most of what we do.

- Now, in this additive framework, it seems obvious that we define the **treatment effect** as

$$Y_1 + \delta - Y_1 = \delta.$$

- Of course, if we can compute this, we're done—no need for randomized experiments, no need for statistics (if our assumption about additive effects holds for our population of interest)!
- Before you get too excited, the general consensus, among statisticians, is that we cannot observe both  $Y_1$  and  $Y_1 + \delta$  because, as the argument goes, we cannot assign the **same** unit to both treatments.
- We avoid the obvious philosophical debate, and simply acknowledge the conservative consensus that we cannot do this. So much for our simple example with one unit.

- Again, see our previous remark (1.1) about the Field of Causal Inference, which regularly does consider such hypothetical (counterfactual) questions as, ‘what if a unit had been assigned to a different group?’

**Example 1.2** (A Slightly Less Simple Situation).

- Now, let’s consider the slightly less simple situation of assigning  $n_1 = 2$  units to treatment level 1 and  $n_2 = 2$  to level 2, ( $T = 2$ ) *See our tables, below.* The particular random assignment that we have to work with will be one of  $\boxed{??}$  possible assignments.
- How do we assess a treatment difference?
- We could take one observation in treatment group 2 and subtract one observation in treatment group 1, i.e.,

$$Y_3 + \delta - Y_1 = 5 - 3 = 2$$

But this does not seem helpful; while we observe  $Y_3 + \delta = 5$ , we don’t know  $\delta$ , so we don’t know  $Y_3$ , so we cannot compute  $Y_3 - Y_1$ , and we cannot get at  $\delta$ . Further, if this difference (2) is somehow large/small, we do not know if it’s due to  $\delta$  or to the particular unit responses  $Y_1$  and  $Y_3$ .

- Perhaps we suspect somehow that using averages is better:

$$\bar{Y}_2 - \bar{Y}_1 = (Y_3 + \delta + Y_4 + \delta)/2 - (Y_1 + Y_2)/2 \quad (1.1)$$

$$= \delta + (Y_3 + Y_4)/2 - (Y_1 + Y_2)/2 \quad (1.2)$$

$$= 2, \quad (1.3)$$

which, again, is not helpful because, again, we cannot compute  $(Y_3 + Y_4)/2$ , and, again, if the difference (2) is large/small, we do not know if it is due to  $\delta$  or to the unit responses.

**Example 1.3** (Hypothetical Replications, Mixing-up and Randomization Distribution).

- Now let's return to our hypothetical replications (assignments) and mixing-up. (*See more tables.*)
- For each assignment, we could (hypothetically) compute

$$\bar{Y}_2 - \bar{Y}_1 = \delta + (Y_{i_1} + Y_{i_2})/2 - (Y_{i_3} + Y_{i_4})/2,$$

(where unit subscripts depend on which assignment we consider).

- We already agreed that we cannot assign the same unit to different treatments, so, for the moment, we can only compute this difference for one replication (the one that we get (2)).
- But, if we assume a value for  $\delta$  then we can compute the differences for the remaining possible assignments. Right? We will *do this with our tables, in class*. How does this help? It may seem strange to assume a value for something that we don't know but want to know.
- Because we randomized units to treatments, we have mixed-up units' confounding variables so that

$$(Y_{i_1} + Y_{i_2})/2 - (Y_{i_3} + Y_{i_4})/2$$

will tend to be the same across assignments. Thus, if our actual observed value, computed from the actual values in our data,

$$\bar{Y}_2 - \bar{Y}_1 = \delta + (Y_3 + Y_4)/2 - (Y_1 + Y_2)/2 = 2,$$

is large/small compared to the other (hypothetical) values,

$$\bar{Y}_2 - \bar{Y}_1 = \delta + (Y_{i_1} + Y_{i_2})/2 - (Y_{i_3} + Y_{i_4})/2,$$

(to be computed shortly), then we tend to think that this because the actual value of  $\delta$  is relatively large/small compared to the assumed value.

- It seems natural to assume  $\delta = 0$  so that large/small value for our observed difference (2), compared to the other hypothetical differences computed under the assumed value of  $\delta = 0$ , tends to make us think that the unknown, actual value of  $\delta$  is larger/smaller than zero, i.e., that the actual value of  $\delta$  is not zero, i.e., that there is a treatment effect!
- (If we had assumed that  $\delta \neq 0$ , we could subtract this value from the treatment group values, for each assignment, including our observed assignment, and, once again, a resulting large/small value of our observed difference compared to the others would tend to indicate a treatment effect.)
- If we haven't already done so by this point, let's complete the tables/computations, in class, according to the above discussion.

- Our Assignment Tables to accompany our examples, above.

First, suppose our actual randomization gives

|       | Unit Response |       |       |       |
|-------|---------------|-------|-------|-------|
| Group | $Y_1$         | $Y_2$ | $Y_3$ | $Y_4$ |
| 1     | 3             | 2     | NA    | NA    |
| 2     | NA            | NA    | 5     | 4     |

- What is the average of the first group?  $\bar{Y}_1 = ???$ .
- Second group ?  $\bar{Y}_2 = ???$ .
- The difference,  $\bar{Y}_2 - \bar{Y}_1 = ???$
- What are all 6 possible arrangements? (we'll circle things, compute things, and make a dot plot in class...got pencil and paper?)

| Group | Unit Response  |                |                |                |
|-------|----------------|----------------|----------------|----------------|
| 1     | $Y_1$          | $Y_2$          | $Y_3$          | $Y_4$          |
| 2     | $Y_1 + \delta$ | $Y_2 + \delta$ | $Y_3 + \delta$ | $Y_4 + \delta$ |

| Group | Unit Response  |                |                |                |
|-------|----------------|----------------|----------------|----------------|
| 1     | $Y_1$          | $Y_2$          | $Y_3$          | $Y_4$          |
| 2     | $Y_1 + \delta$ | $Y_2 + \delta$ | $Y_3 + \delta$ | $Y_4 + \delta$ |

| Group | Unit Response  |                |                |                |
|-------|----------------|----------------|----------------|----------------|
| 1     | $Y_1$          | $Y_2$          | $Y_3$          | $Y_4$          |
| 2     | $Y_1 + \delta$ | $Y_2 + \delta$ | $Y_3 + \delta$ | $Y_4 + \delta$ |

| Group | Unit Response  |                |                |                |
|-------|----------------|----------------|----------------|----------------|
| 1     | $Y_1$          | $Y_2$          | $Y_3$          | $Y_4$          |
| 2     | $Y_1 + \delta$ | $Y_2 + \delta$ | $Y_3 + \delta$ | $Y_4 + \delta$ |

| Group | Unit Response  |                |                |                |
|-------|----------------|----------------|----------------|----------------|
| 1     | $Y_1$          | $Y_2$          | $Y_3$          | $Y_4$          |
| 2     | $Y_1 + \delta$ | $Y_2 + \delta$ | $Y_3 + \delta$ | $Y_4 + \delta$ |

| Group | Unit Response  |                |                |                |
|-------|----------------|----------------|----------------|----------------|
| 1     | $Y_1$          | $Y_2$          | $Y_3$          | $Y_4$          |
| 2     | $Y_1 + \delta$ | $Y_2 + \delta$ | $Y_3 + \delta$ | $Y_4 + \delta$ |

- Can we compute  $\bar{Y}_2 - \bar{Y}_1$  for each of the possible arrangements? (Recall that we said that we cannot get both,  $Y_i$  and  $Y_i + \delta$  for the *same* unit, *i.*)
- What if we assume  $\delta = 0$  (just an assumption, not necessarily true, of course)? Under this assumption, now can we compute  $\bar{Y}_2 - \bar{Y}_1$  for each of the possible arrangements? We'll do this **in class**.
- The 6 values of  $\bar{Y}_2 - \bar{Y}_1$  (**to be**) computed in the previous bullet, under the assumption of  $\delta = 0$ , constitute the **randomization distribution** of  $\bar{Y}_2 - \bar{Y}_1$ . **We will draw this on the board.**
- **p-value.** To somehow reach some closure on our introduction to the randomization distribution, let's compute a p-value (**in class**). We give a formal definition, below.

**Remark 1.7** (Example Summary).

- We **summarize and formalize** a few things in the current example before giving another example of a randomization distribution and its use in statistical inference.
- **Null Hypothesis.** A key assumption, above, which allowed us to proceed, was that  $\delta = \delta_0$  where  $\delta_0$  is some specified value, commonly zero, as above. We may write, e.g.,

$$H_0 : \delta = 0,$$

and call this a null hypothesis in the sense that it says there is (we assume) no treatment effect (treatment is “null” so-to-speak). (To be sure, we assumed a value of zero to construct the randomization distribution, but, as mentioned, above, we can use this randomization distribution to test any “null” hypothesized value, zero or otherwise.)

- **Alternative Hypothesis.** Of course, we don’t know what  $\delta$  is. Generally, we expect  $\delta \neq \delta_0$  (two-sided alternative) or  $\delta > \delta_0$  or  $\delta < \delta_0$  (one-sided alternatives). We refer to these as alternative hypotheses, and we may write, e.g.,

$$H_a : \delta > 0.$$

- What’s the “Difference”? Generally speaking, the null hypothesis is that outcomes do not “differ” between treatment groups or that there is no association between outcomes and treatment group status. (Similarly for the alternative hypothesis.) Because we **assumed** a particular form of difference (**additive effect**), we stated the null hypothesis in terms of this assumption, in terms of  $\delta$ , which seems relatively straightforward, doesn’t it? Again, additivity effects are very common.

- Our example illustrates a few basic concepts that we define more formally, below.

**Definition 1.18** (Parameter).

- *A parameter is an unknown constant associated with a model used for answering questions of interest.*
- *Usually, questions (and answers) are phrased in terms of parameters (“is  $\delta = 0?$ ,” in the case of a test, or “what is  $\delta?$ ,” in the case of an interval) or functions of parameters.*

**Definition 1.19** (Statistic). *Some function of data that does not depend on unknown parameters.*

**Definition 1.20** (Probability Distribution). *A description of the pattern of values (i.e., relative frequency) of a random variable.*

**Definition 1.21** (Estimator/Estimate). *When a statistic is used to infer about a parameter, we often call the statistic an estimator (before plugging in values) or an estimate (after plugging in values).*

**Definition 1.22** (p-value). *The probability of a randomization resulting in a test statistic value as extreme as or more extreme than the value observed for the actual randomization obtained in a randomized experiment. (We tailor the definition here to the context at hand, and will offer a similar, more typical definition, later.)*

- In our example, the treatment effect,  $\delta$ , is a **parameter**, and we may view it as the (unknown) “location” of the distribution of  $\bar{Y}_2 - \bar{Y}_1$ , which is a **statistic**.
- Note that, before we plug in our actual data values,  $\bar{Y}_2 - \bar{Y}_1$  is uncertain, i.e., it’s an example of what we defined to be a random variable, whose uncertain values are characterized by its (randomization) distribution (arising from the randomization mechanism) that we used to answer the question of whether  $\delta = \delta_0$  or not (or could have used to construct an interval). In particular, we used the (randomization) **distribution** of  $\bar{Y}_2 - \bar{Y}_1$  under the assumption that  $\delta = 0$ .
- The statistic,  $\bar{Y}_2 - \bar{Y}_1$ , is also an **estimator/estimate** because we may use it to infer about (“estimate”) the parameter,  $\delta$  (why? how?).
- We also may call  $\bar{Y}_2 - \bar{Y}_1$  a **test statistic**, because we used it to “test” whether  $\delta = 0$  or not.
- In our illustration of the randomization distribution, we computed a **p-value** of ???

**Remark 1.8** (Interpreting the p-value).

- *Probability of observing data (or a statistic) at least as extreme as*

*those (the one) actually observed, given that the null hypothesis is true.*

- *A measure of how plausible our data (or statistic) is to have arisen from a particular distribution (the distribution under the null hypothesis).*
- *The proportion of times (relative frequency) that we would observe data (or a statistic) at least as extreme as those (the one) we actually did observe, in a large number of **hypothetical replications** of our test procedure (same units, same treatments, many different random assignments (in the current context of randomized experiments), same null hypothesis, same other assumptions)*
- *Generally NOT the probability that the null hypothesis is true (though, there exist particular instances of a particular statistical framework (Bayesian statistics) wherein p-values do coincide exactly with probabilities that (particular) null hypotheses are true...)*

#### Example 1.4 (Motivation and Creativity Randomized Experiment).

- *Here, we use the Motivation and Creativity Experiment of [RS13, Section 1.1.1 and 1.3] to illustrate the use of the randomization distribution to characterize the uncertainty in the difference of average creative ability scores (a statistic) between two treatment groups of children.*
- *More explanation in class. There is no need for the textbook by [RS13].*

- We're about to confront R for the first time, so we can take some time here, if you want to.
- Data are available from the add-on R package called **Sleuth3**, which you likely must download before you use the `library(Sleuth3)` command.
- I will give you code ("code chunks": we'll learn to love code chunks) to reproduce nearly all statistical results you see in these notes. Attending lectures may be very informative!!!
- Don't worry if R initially looks foreign. It takes time to learn R. We'll learn mostly by example in this class, with much in-class discussion. Again, come to class!
- To R!

```
> library(Sleuth3) ## likely need to download first
> case0101.df<- case0101 ## not necessary
> head(case0101.df ,n=3); tail(case0101.df ,n=3)

  Score Treatment
1    5.0   Extrinsic
2    5.4   Extrinsic
3    6.1   Extrinsic
  Score Treatment
45   24.3  Intrinsic
46   26.7  Intrinsic
47   29.7  Intrinsic

> attach(case0101.df)
> tapply(Score,Treatment,length)

Extrinsic Intrinsic
      23          24
```

```
> (mscore<- tapply(Score,Treatment,mean))  
  
Extrinsic Intrinsic  
15.73913 19.88333  
  
> tapply(Score,Treatment,sd)  
  
Extrinsic Intrinsic  
5.252596 4.439513  
  
> diff(mscore)  
  
Intrinsic  
4.144203  
  
> detach(case0101.df)  
> detach(package:Sleuth3)
```

- Now, we are about to create the (approximate (**why?**)) randomization distribution in R (not by hand as in our previous example!).
- Assuming that there is no difference among the two treatment groups, i.e., that the treatment effect is zero, then, under this null assumption, i.e., **null hypothesis**,  $\delta = 0$ , each subject *would have* received the same score no matter which treatment the subject is assigned to, thus we can **mix-up** the values among different arrangements, as we did in our previous example, to get the (approximate) randomization distribution of the (test) **statistic**,  $\bar{Y}_2 - \bar{Y}_1$  (Intrinsic - Extrinsic).
- If, in reality, treatments do affect creativity scores (**alternative hypothesis**,  $\delta \neq 0$ ), then the actual difference in average creativity scores computed from our data, as actually observed (i.e., not the re-randomized versions), will tend to be more extreme than expected under the randomization distribution.

- Lets compute the (approximate) randomization distribution and (approximate) p-value.

```

> attach(case0101.df) ## from previous chunk
> M<- 10000 ## # of Monte Carlo (MC) assignments
> dscore<- vector("numeric", length=M)
> dscore[1]<- diff(tapply(Score, Treatment, mean))
> score<- Score
> set.seed(8675309)
> for (i in 2:10000){
+   score<- sample(score)
+   mscore<- tapply(score, Treatment, mean)
+   dscore[i]<- diff(mscore)
+ }
> detach(case0101.df)
> (pval<- sum(abs(dscore) >= abs(dscore[1]))/10000)

[1] 0.0048

> mean(abs(dscore) >= abs(dscore[1]))

[1] 0.0048

> choose(47,23) ## !...hence MC approximation

[1] 1.61238e+13

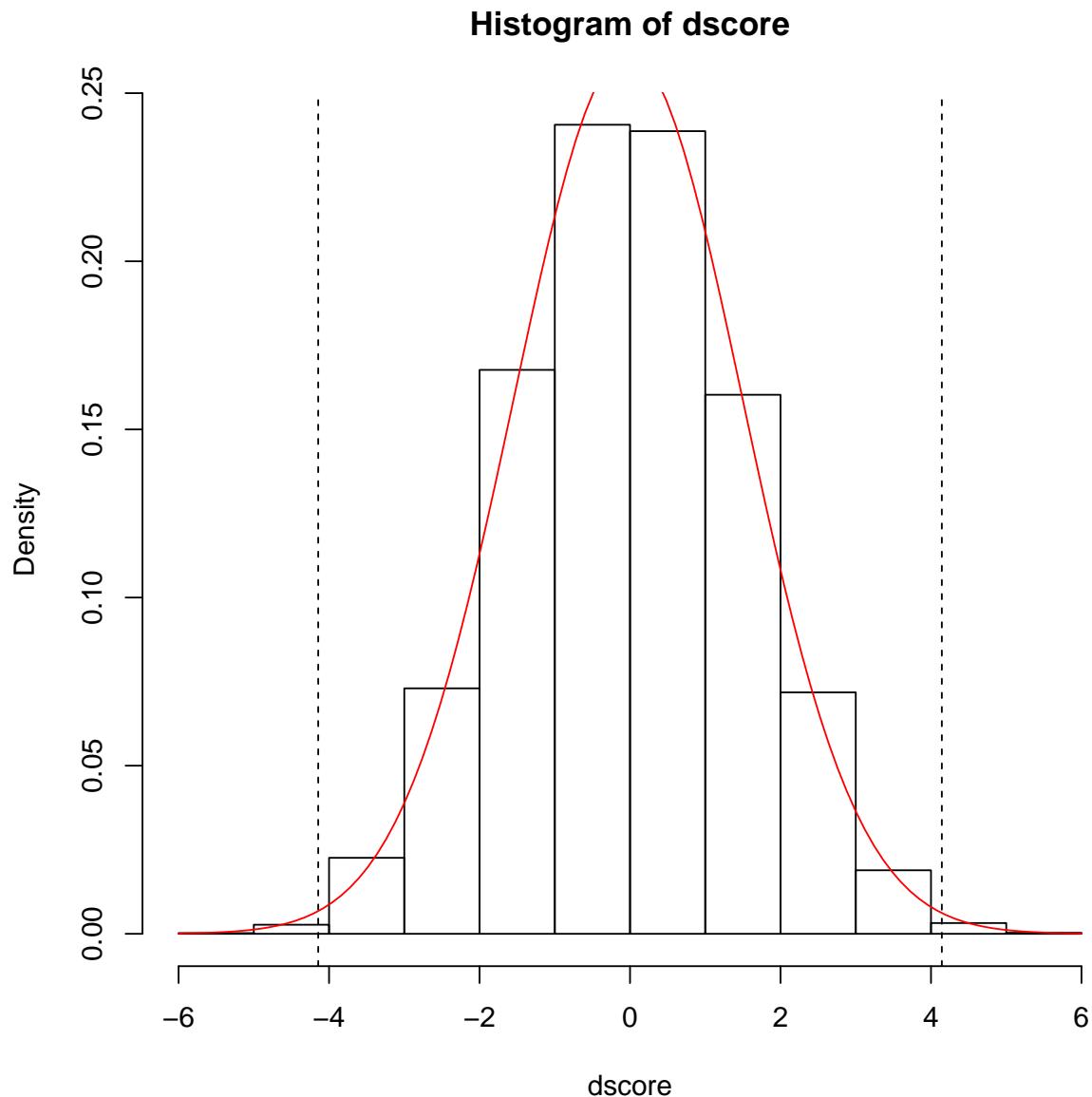
```

- So, the observed (approximate) two-sided p-value is 0.0048.
- That means *if* the null hypothesis of no treatment effect is true, then the probability of observing a difference of average creativity scores

at least as extreme as the one we actually observed (4.1442) is about 0.0048, or 48 in 10000.

- Of course, an extreme value (i.e., small p-value) may also be due to our assumption of no treatment effect being false, in which case we may choose some **significance level** (e.g., 0.05) at or beyond which we formally reject the null hypothesis.
- The p-value is illustrated as the area in the tails (indicated by the vertical lines) of the randomization distribution histogram in the nearby figure.

```
> hist(dscore, freq=FALSE)
> abline(v=c(-1,1)*dscore[1], lty=2)
> curve(dnorm(x, m=mean(dscore), s=sd(dscore)),
+         add=TRUE, col="red") ## BTW...
```



Some questions/review:

- Null and alternative hypotheses?
- What do we mean by “difference” between groups?
- What is the scope of inference?

- Can we (statistically) infer a cause-effect relationship?
- Can we (statistically) infer to a population? (perhaps a bit ahead of ourselves with this question)
- Which box (Figure 1.1) does this example fit into?
- How do you interpret the p-value?
- Conclusion?

- MC sample approximation
- Normal approximation...
- Histogram approximation

- All of this talk about randomized experiments and causal inference may suggest that observational studies, which, by our own definition, above, are not randomized, are not important.
- See [RS13, Page 7] for comments on the value of observational studies, or see our similar comments about this, below.

## 1.5 Sampling Distribution

- Analogous to the randomization mechanism of randomized experiments and its associated reference (null) distribution (randomization distribution) arising from the notion of hypothetical replication of the

mechanism, we also have the mechanism of random sampling, which also leads to the notion of hypothetical replications and a reference distribution for a statistic.

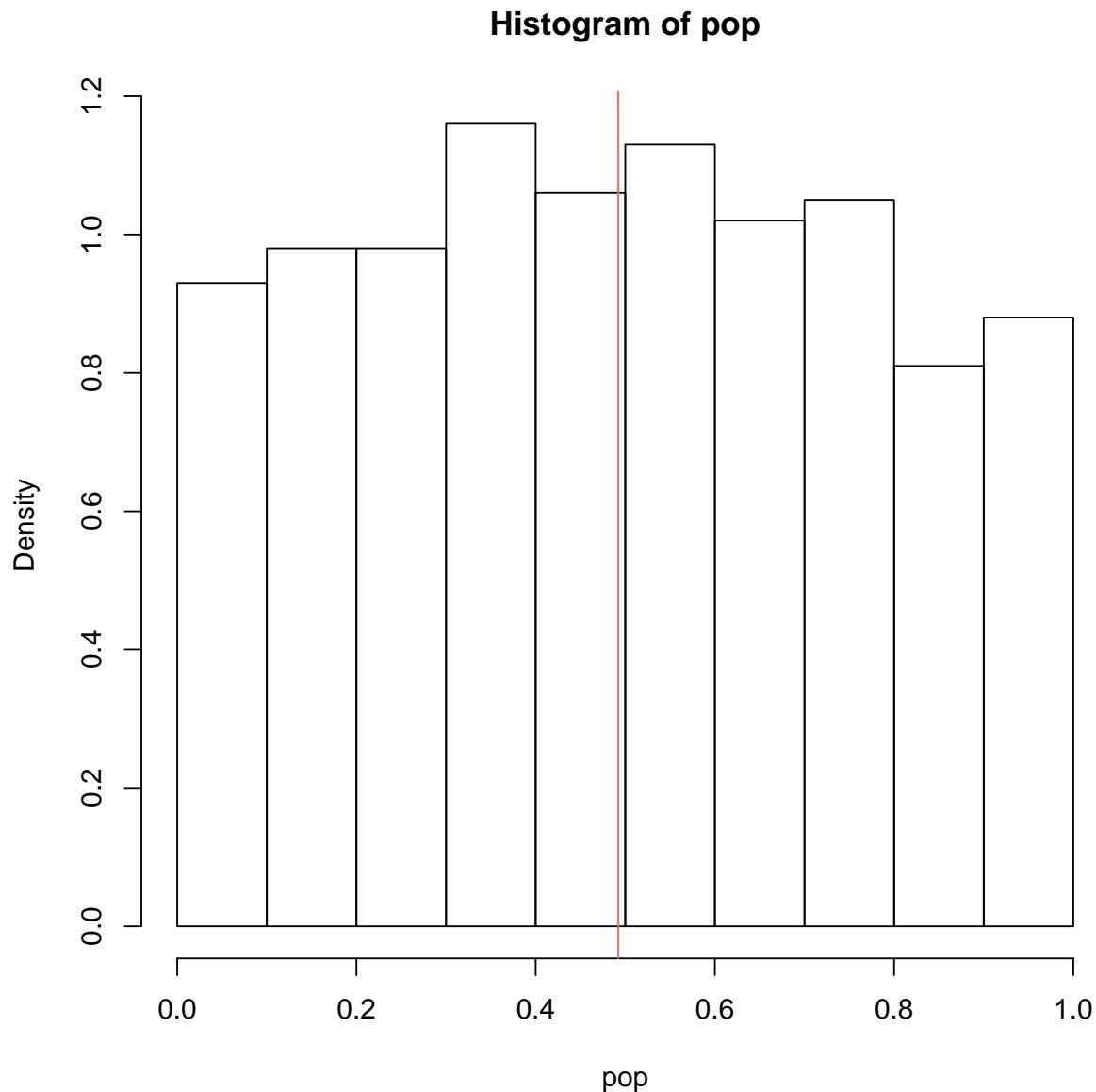
- For random sampling, we consider (hypothetically) **all possible random samples** (of whatever size we are considering, say, generically,  $n$ ) from the same population.
- Then, we consider all corresponding values of a (test) **statistic**, e.g., sample average if, say, we want to infer about the population mean (like in “STAT 101”).
- A histogram of these replicated test statistics would represent the **sampling distribution** for that statistic.
- We could then use the sampling distribution to test hypotheses, etc., similar to our discussion of the randomization distribution, above.
- But, of course, if we could obtain all possible samples of size  $n$  from the population, we would have access to the entire population and would have its mean and its variance and all other properties; there would be no need for statistics!
- So, how do we proceed?
- In the random sampling case, we appeal to **theory**. In particular, the **Central Limit Theorem** tells us that, practically speaking, no matter what population we consider, as long as our (random) sample size,  $n$ , is “large enough”, then **sample averages are approximately normally distributed**. (Assume that sample averages are appropriate statistics/estimators, as if we want to estimate population means.)
- Or—this may sound familiar—we might **assume** at the outset that our population distribution is normal, which is an obvious approximation. (Why?) (We might offer plots, e.g., histograms, and summary statistics to help build support for this assumption.)

- So, we appeal to theory and/or assumptions to get us to an **approximate sampling distribution**, e.g., the **typical normal/t based procedures** for a single random sample that you've seen before in a previous stats course ("STAT 101").

### Example 1.5 (Illustrating a Sampling Distribution via CLT).

- To illustrate, we generate  $N = 1000$  values from a uniform distribution to mimic the responses from a **population** of  $N$  units.
- Further, assume we are interested in inferring the **mean of the population**, which we know in this example, to be  $\mu = \sum_{i=1}^N Y_i/N = ???$ , where, like in our randomized experiment,  $Y_i$  denotes the response of unit  $i$ , now for all units in the population.
- We also compute the **population standard deviation**.
- (Of course, we usually cannot compute the population mean/sd in practice.)

```
> ## CLT
> N<- 1000 ## Population size
> set.seed(20500 + 5150 + 24601)
> pop<- runif(n=N,min=0,max=1) ## ~ superpopulation mean 0.5
> (mu<- mean(pop)) ## pop mean
[1] 0.4921489
> (sigma<- sd(pop)) ## pop std dev.
[1] 0.2784678
> hist(pop, prob=TRUE)
> abline(v=mu, col="tomato")
```



- (E.g., cont'd) Now, proceeding as in practice, without knowledge of the population, we mimic a simple random sample of size  $n = 30$  and compute the average,  $\bar{Y}$ , as a **statistic**.

```
> n<- 30 ## sample size
> oursampley<- sample(x=pop, size=n, replace=FALSE)
> (ourybar<- mean(oursampley))
```

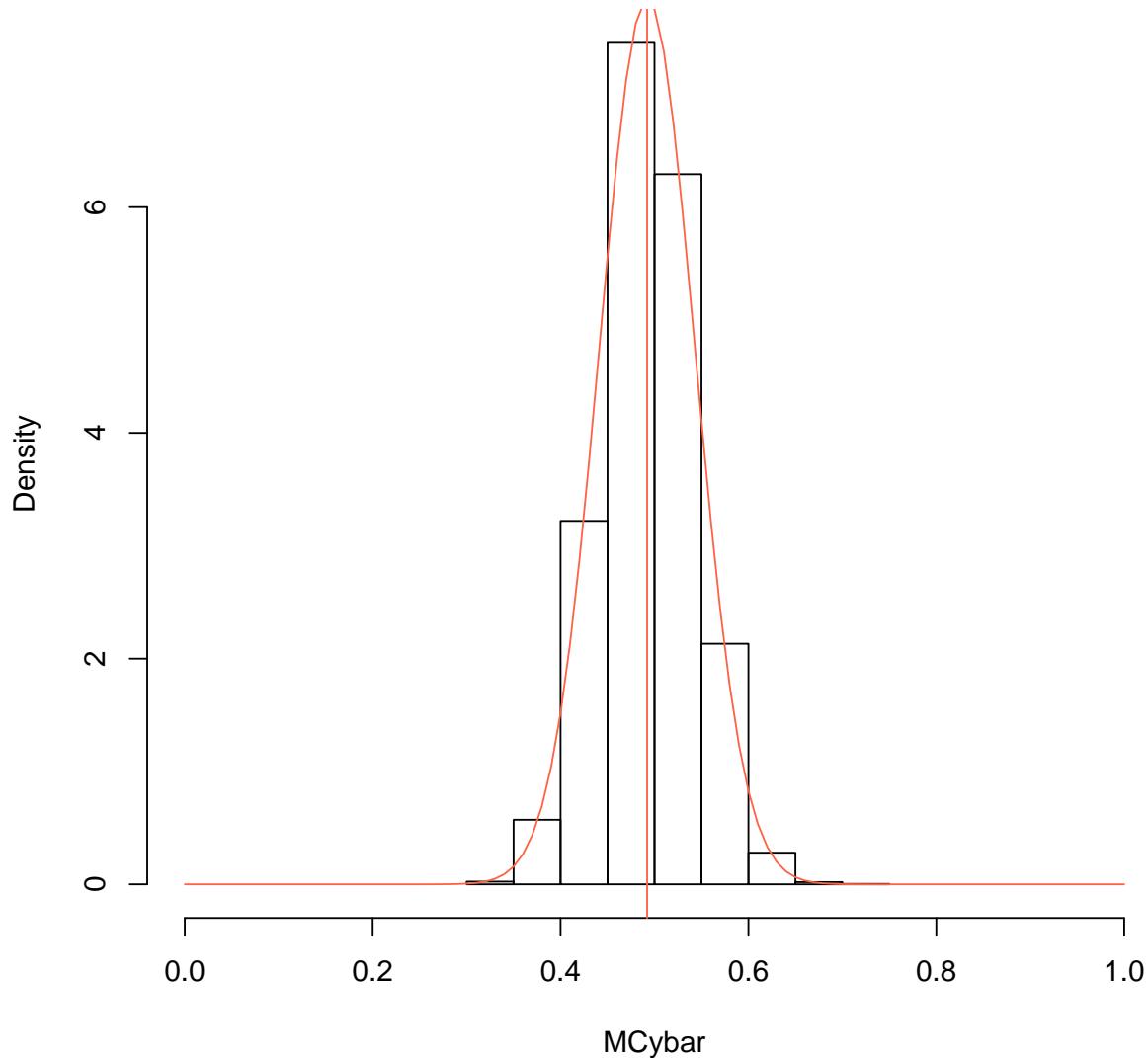
```
[1] 0.4357687
```

- (E.g. cont'd) How do we use this to infer about the population mean? Again, if we could compute our statistic for all possible random samples, then we would know the population, and would have no need for Statistics. (Curb your excitement.) In practice, we obtain only one sample and (in this example), one statistic.
- We said that we appeal to theory/assumptions. In particular, we appeal to the **Central Limit Theorem** in this example, which says, loosely, that sample means are approximately normal (despite the fact that the population distribution, as depicted by a histogram, does not appear normal).
- Because we know the population in this example, we could compute our statistic for all possible random samples, in principle, so we can compare the actual sampling distribution to the one provided by the CLT, just to convince you that the CLT gives a good approximation to the actual sampling distribution.
- However, in our example, there are far too many random samples (see code) to enumerate and compare a statistic for to obtain the actual sampling distribution. Instead, we obtain a Monte Carlo (MC) sample of, size, say,  $M = 5000$ , from all possible random samples of size  $n$ , compute the average for each of the  $M$  samples, then display the resulting histogram of  $M$  averages as an approximation to the unobtainable sampling distribution.
- What do we notice? Does theory seem to be operating here? You may recall the CLT tells us that  $\bar{Y} \sim N(\mu, \sigma^2/n)$ , at least approximately.

```
> ## How many possible random samples?  
> choose(N,n) ## far too many to enumerate
```

```
[1] 2.429608e+57
```

```
> ## So, we obtain  $M$  MC samples of samples of size  $n$  as approximation.  
> M<- 5000  
> MCybar<- vector("numeric",length=0)  
> for (i in 1:M) {  
+   MCsampley<- sample(x=pop,size=n, replace=FALSE)  
+   MCybar<- c(MCybar, mean(MCsampley))  
+ }  
> hist(MCybar, prob=TRUE, xlim=c(0,1))  
>  
> ## Compare to CLT  
> curve(dnorm(x,mean=mu, sd=sigma/sqrt(n)), from=0, to=1,  
+       add=TRUE, col="tomato")  
> abline(v=mu, col="tomato")
```

**Histogram of MCybar**

- (E.g. cont'd) Of course, we used the population mean and standard deviation,  $\mu$  and  $\sigma$ , to illustrate that the CLT is working as an approximation to the sampling distribution of  $\bar{Y}$ .
- In practice, again, we don't know  $\mu$  and  $\sigma$ , but, still, the CLT says  $\bar{Y}$  is approximately normal. In this case, we play a similar game to our randomization experiment above and assume that we know  $\mu$ , which we

formulate in a **null hypothesis**,

$$H_0 : \mu = 0.4,$$

where we choose 0.4 merely for illustration.

- Again, similar to our randomization distribution discussion, the idea is that, if our actual, observed statistic,  $\bar{Y}$  is large/small relative to those values in the null distribution, then we **tend** to think that this is due to the actual value of  $\mu$  being larger/smaller than hypothesized under the null distribution.
- Still, in our current example, we don't know the population standard deviation,  $\sigma$ . Sparing you the remaining details (you should have at least heard of them before), we end up using the **Student's *t* distribution** (with  $n - 1 = 29$  degrees of freedom) as the (approximate) sampling distribution of the statistic,

$$\frac{\bar{Y} - 0.4}{\frac{s}{\sqrt{n}}},$$

for which our observed value is **???**.

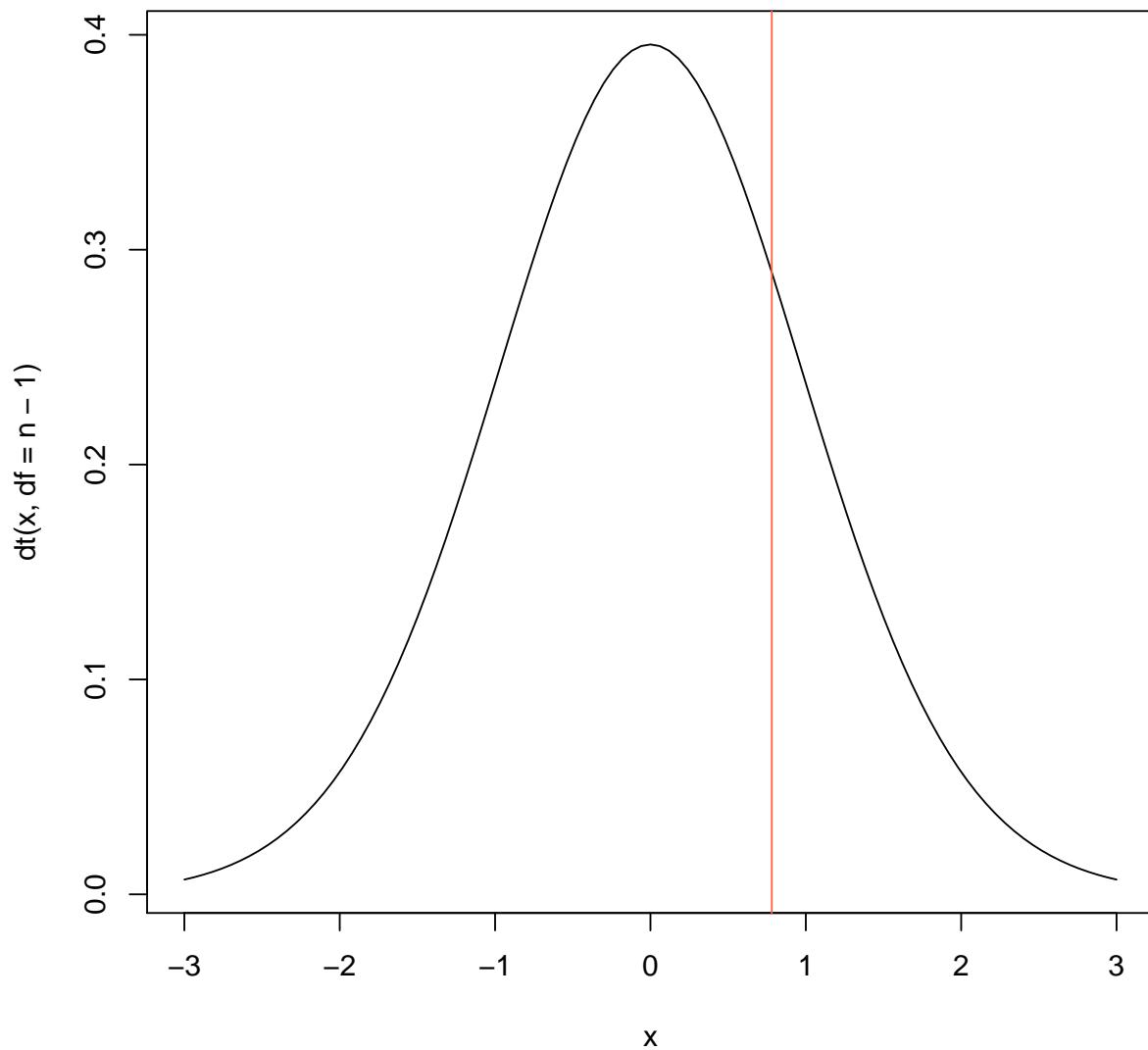
```
> mu0<- 0.4
> (s<- sd(oursampley))

[1] 0.250881

> (ourtstat<- (ourybar - mu0) / (s / sqrt(n)))

[1] 0.7809005

> curve(dt(x,df=n-1), from=-3, to=3)
> abline(v=ourtstat, col="tomato")
```



```
> (pval<- 2 * (1 - pt(ourtstat, df=n-1)))  
[1] 0.4411894
```

- In practice, we often go directly to mathematical approximations of randomization distributions or sampling distributions, which are, of

course, typically normal distributions—as illustrated in plots of normal distributions overlain onto histograms, above—or related distributions, like a  $\chi^2$  or  $t$  or  $F$ .

- In this case, these approximating distributions can be thought to represent a **superpopulation** of infinite size ( $\infty$ ), instead of the actual population size,  $N$ , or the sample size,  $n$ , ([Wak13, p. 3]). Incidentally, this is likely why some statisticians often like to refer to a population of values rather than a population of units, because they skip the actual population to get to the approximating theoretical probability distribution (superpopulation) of numbers
- But, despite this nearly universal use of approximating distributions, (frequentist, at least) statisticians still appeal to the notion of **hypothetical replications** associated with some finite set of possibilities.
- For example, consider the frequentist interpretation of the p-value, which is the probability, under the null distribution, of observing a test statistic at least as extreme as the one we observed: “if we repeat our sampling (and/or randomization) a large number of times, from the same population(s) (and/or experimental setup), and compute a test statistic for each repeated sample, then the p-value is approximated by the proportion of test statistics that are as extreme or more extreme than the one we actually computed.” In other words, the p-value is the long-run (as the number of replications gets large) relative frequency (proportion) of such extreme test statistics.
- Thus, another take home message is the fundamental notion of hypothetical replications, or **long-run relative frequency interpretation** of probability of frequentist statistics. Indeed, for frequentist statisticians, probability in general, not just for p-values, is interpreted as a long-run relative frequency of some event.
- Now that we have some idea of how the normal distribution works as an approximation to the fundamental distributions arising from

random mechanisms, we will tend to start with the normal distribution (or related distribution) as an approximation for our data or statistics.

- We should mention that there is an entire enterprise of finite population ( $N$ ) statistics (survey sampling), but we leave this to other courses.

## 1.6 Notes on Observational Studies

- Inference from **observational data** (no randomization of units to treatments) is limited.
- Without randomization of units, we are not on firm ground with regard to causality.
- Further, if observational data are not based on a random sample, then we are not on firm ground with regard to inference about a larger population of interest.
- Still, **observational studies may be compelling** for other reasons.
- For example, if a **drug appears to have a different effect for men vs. women**, this information may lead doctors do prescribe the drug differently for men and women, regardless of underlying causal mechanisms. (Note that we cannot randomly assign people to gender.)
- As another example, consider **smoking and cancer**. If we observe elevated cancer rates among smokers across practically all potential values of confounding variables (gender, age, ethnicity, diet, etc., etc.), then this suggests that we would see the same results regardless of mixing-up (randomizing) confounding variables, hence providing very compelling causal evidence. (Though it often takes a long

time—and at great loss in case of smoking and cancer—to make observations over practically all conditions.)

- Also, observational studies may compel us to perform **follow-up** randomized experiments, suggest **hypotheses** to explore or provide evidence to support/establish **theory**.
- And, as we mentioned, if we are willing or are compelled to make further assumptions about causality, then we might explore the field of **Causal Inference** (remark 1.1) can make causal statements from observational data.

## 1.7 Unobserved Covariates and “Randomness”

Sometimes, we do not have a random mechanism that generates a probability distribution to facilitate inference. That is, our data may not arise from random sampling of units from a population and/or random assignment of units to treatments. Still, in such cases, we very often find ourselves using the normal distribution to characterize errors.

[Wak13, §1.4] discusses how the omission of (the sum of many small) unmeasured covariates (confounding variables) is manifested, via the CLT, into approximately normal errors. This notion of errors in a regression model arising from unmeasured covariates is also reflected in the use of (explicit, measured) covariates in regression models to “explain” error, but your author cautions against the explanation that covariates cause outcomes, in (nonrandomized) observational studies by illustrating how unmeasured covariates (confounders) can change the meaning of parameters.

Thus, we may argue that the omission of many small unmeasured covariates (perturbations) tends to support the use of a normal distribution for regression errors, but observational studies are still limited in their ability to infer a causal effect between outcome and covariate.

## 1.8 Subjective Probability

The other main philosophical school of thought on probability and statistics, aside from frequentism, is subjective probability and **Bayesian** statistics. The author of your textbook and this course make a good effort at a pragmatic approach to statistics, Bayesian or frequentist. We will learn more about both as we go. For now, we simply say that the subjective view of probability and (subjective) Bayesian statistics view probability as degrees of belief about uncertain quantities ([Wak13, §1.5]), which may be viewed to subsume the frequentist view inasmuch as the frequentist view may motivate belief. It's been said that, because each of us has her/his own subjective belief about the same event, **probabilities do not exist!**

## 1.9 Summary

Random mechanisms mix-up confounding variables and lead to the notions of hypothetical replications and the long-run relative frequency interpretation of probability that underlies frequentist statistics. We will consider both frequentist and Bayesian statistical methods.

For **randomized experiments**, the mix-up **tends** to give (leads us to **expect**) no systematic effect (bias) on outcomes among different treatment levels due to a different composition units' confounding variables among the levels. Thus, upon observing systematic differences in outcomes among treatment levels, we tend to think these are due to the treatment variable; i.e., we may conclude that treatment has a causal effect on outcome. Still, our particular random assignment of units may, by chance, consist of a such a particular composition of confounding variables that do exhibit systematic differences in the outcomes across treatment levels, and we won't know if this is the case, in practice. However, in the absense of a treatment effect, the null hypothesis and associated null distribution—the randomization distribution of a randomized experiment—allows us to compute the probability of falsely rejecting the null hypothesis (probability of a type I error). Thus, we can tame the probability of making such

a mistake by choosing a small type I error probability beyond which we are comfortable concluding a causal treatment effect, a scope of inference toward a mechanistic relationship beyond mere association of outcomes and treatments due to shared association with unobserved confounding variables.

For **random sampling** from a population, the mix-up leads us to **expect** no systematic effect (bias) in the sample due to a different composition of units in the sample (“biased sample”) compared to the population, i.e., it **tends** to give us an unbiased sample. Thus, upon observing a sample result, we tend to think it applies also to the population, broadening our scope of inference beyond the sample to a larger group of units of interest. Still, by chance, we may obtain a particular sample of units whose composition of confounding variables cause us to make incorrect conclusions about the population, and we won’t know if this is the case, in practice. However, the sampling distribution (null distribution) allows us to tame the probability of making a false claim (type I error) about the population, and we may choose a small type I error that make us comfortable when inferring that sample results apply to populations.

In either case, probabilities are computed as proportions—or **relative frequency**—of (hypothetical) replications, thus giving rise to the frequentist interpretation of probability as the **long-run relative frequency** or proportion of outcomes in a large number of repeated experiments or samples. And, in either case, these frequencies are very often **approximated by a mathematical formula of probability** (e.g., normal, chi-square, t, F, i.e., “STAT 101” distributions) to which is attached the long-run relative frequency interpretation (for frequentists, at least).

Use of the normal distribution in regression is often argued because of the CLT-like effect of many, small unmeasured covariates, though causal interpretation of model parameters are limited (without further assumptions).

Probability may also be viewed as a subjective degree of believe about the uncertainty of unknown quantities (either random variables or parameters), as in **(subjective) Bayesian statistics**.

# Lecture 2

## Motivating Examples and a Course Outline

### Contents

---

|        |   |    |
|--------|---|----|
| 2.1    | Dental Growth . . . . .                                   | 49 |
| 2.2    | Smoking and Forced Air Expiratory Volume (FEV1) . . . . . | 50 |
| 2.3    | Outcome After Head Injury . . . . .                       | 52 |
| 2.4    | Contraception Drug . . . . .                              | 55 |
| 2.5    | Seizure Data . . . . .                                    | 57 |
| 2.6    | Lung Cancer & Radon . . . . .                             | 61 |
| 2.7    | Aircraft Fastener Data . . . . .                          | 66 |
| 2.8    | Cardiac Failure & Cadralazine Concentration . . . . .     | 71 |
| 2.9    | Pharmacokinetics of Theophylline Data . . . . .           | 74 |
| 2.10   | Course/Textbook Outline . . . . .                         | 77 |
| 2.10.1 | Summary of This Course . . . . .                          | 80 |

---

*Main Objectives:*

- Preview types of data and their modeling in INF 511 (fall) and INF 512 (spring)
- Understand the “big picture” of the course(s) and our textbook in the context of these examples

---

$\mathcal{O}$

*Additional Reading:*

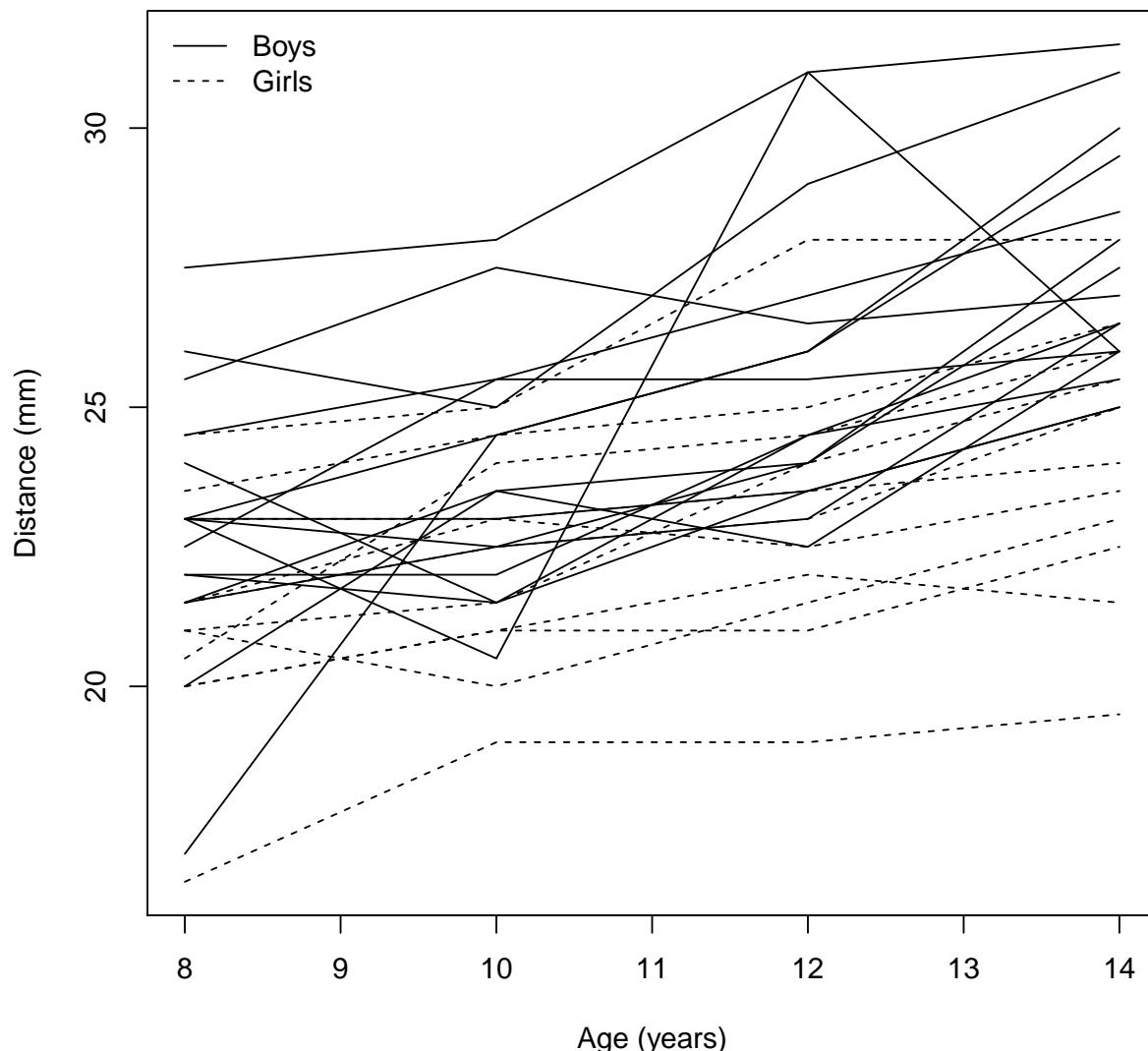
[Wak13, §1.3]

—  
 $\mathcal{R}$

## 2.1 Dental Growth

These data give distance,  $y_{ij}$  (mm), between the center of the pituitary gland to the pterygo-maxillary fissure for observation  $j$  of individual  $i$  at age  $t_{ij}$ , with further categorization by gender,  $G_i$  (0/1, boy/girl).

The following reproduces [Wak13, Fig. 1.8].



One of the models considered later for these data is

$$y_{ij} = (\beta_0 + \beta_1 G_i + b_{i0}) + (\beta_2 + \beta_3 G_i + b_{i1}) t_{ij} + \delta_{ij} + \epsilon_{ij},$$

where  $b_{i0}$  and  $b_{i1}$  are individual random intercept and slope effects, and  $\delta_{ij}$  are random effects to account for residual temporal correlation across time (within individual  $i$ ). See [Wak13, p. 373] (no temporal effects, however,...and where's  $\beta_3$ ?...).

### Some Considerations for these Dental Growth Data:

1. Scope of inference?
2. Aims: (a) predict growth for individuals (“individual” inference or “subject specific” inference or “conditional” inference); (b) estimate average growth for boys, for girls or both (“population averaged” inference or “marginal” inference).
3. Chapter 8 Linear Models (dependent data). LMM. GEE.

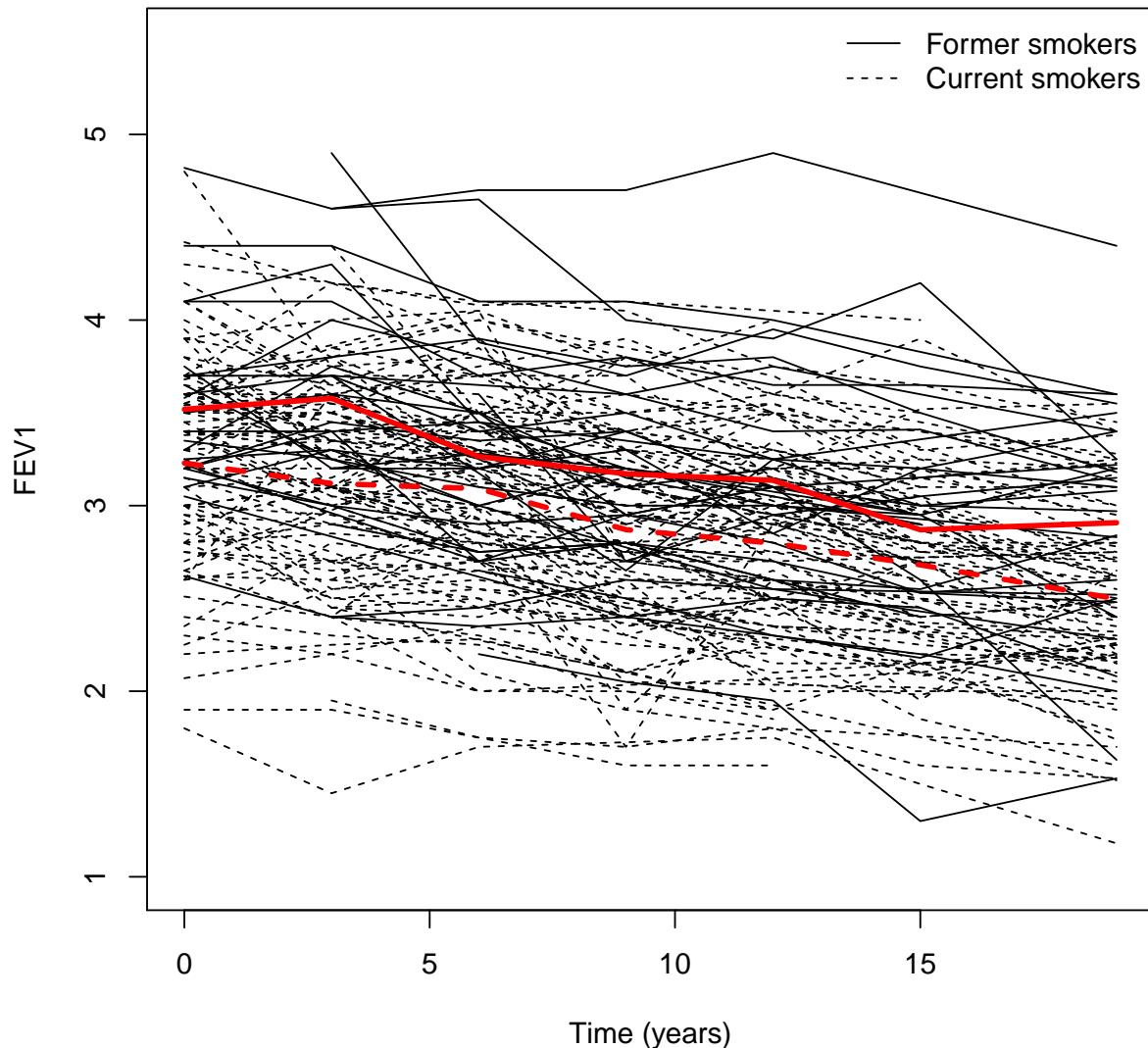
## 2.2 Smoking and Forced Air Expiratory Volume (FEV1)

This data is similar to the dental growth data, but with more subjects to better illustrate model checking (later). The factor here is smoking [ $S_i$ ] (0/1 former smoker/current smoker), not gender, and the response, [ $y_{ij}$ ] is FEV1, a measure of respiratory function.

The following essentially reproduces [Wak13, Tab. 8.4, p.408], averaging the FEV1 over individuals at each time for each smoking group.

|      | smoker | 0    | 1    |
|------|--------|------|------|
| time |        |      |      |
| 0    |        | 3.52 | 3.23 |
| 3    |        | 3.58 | 3.12 |
| 6    |        | 3.26 | 3.09 |
| 9    |        | 3.17 | 2.87 |
| 12   |        | 3.14 | 2.80 |
| 15   |        | 2.87 | 2.68 |
| 19   |        | 2.91 | 2.50 |

The following plot essentially reproduces [Wak13, Figs. 8.5 & 8.6, pp. 408-9].



One of the models considered later for these data is

$$y_{ij} = (\beta_0 + \beta_1 S_i + b_i) + (\beta_2 + \beta_3 S_i)t_{ij} + \epsilon_{ij},$$

where  $b_i$  are individual intercept effects. (See [Wak13, p. 409].) Compared to the dental growth data model, suggested previously, we do not have random slope effects or temporal random effects; keep in mind, this is just one

model, among others, used to illustrate model checking, later, and to motivate discussion, not necessarily to illustrate “the correct” model for these data.

### Some Considerations for these FEV1 Smoking Data:

1. Again, these data are used to illustrate model checking because the dental data have too few observations, but, otherwise, considerations for these data are similar to the dental data.
2. Scope of inference?
3. Chapter 8 Linear Models (dependent data). LMM. GEE.

## 2.3 Outcome After Head Injury

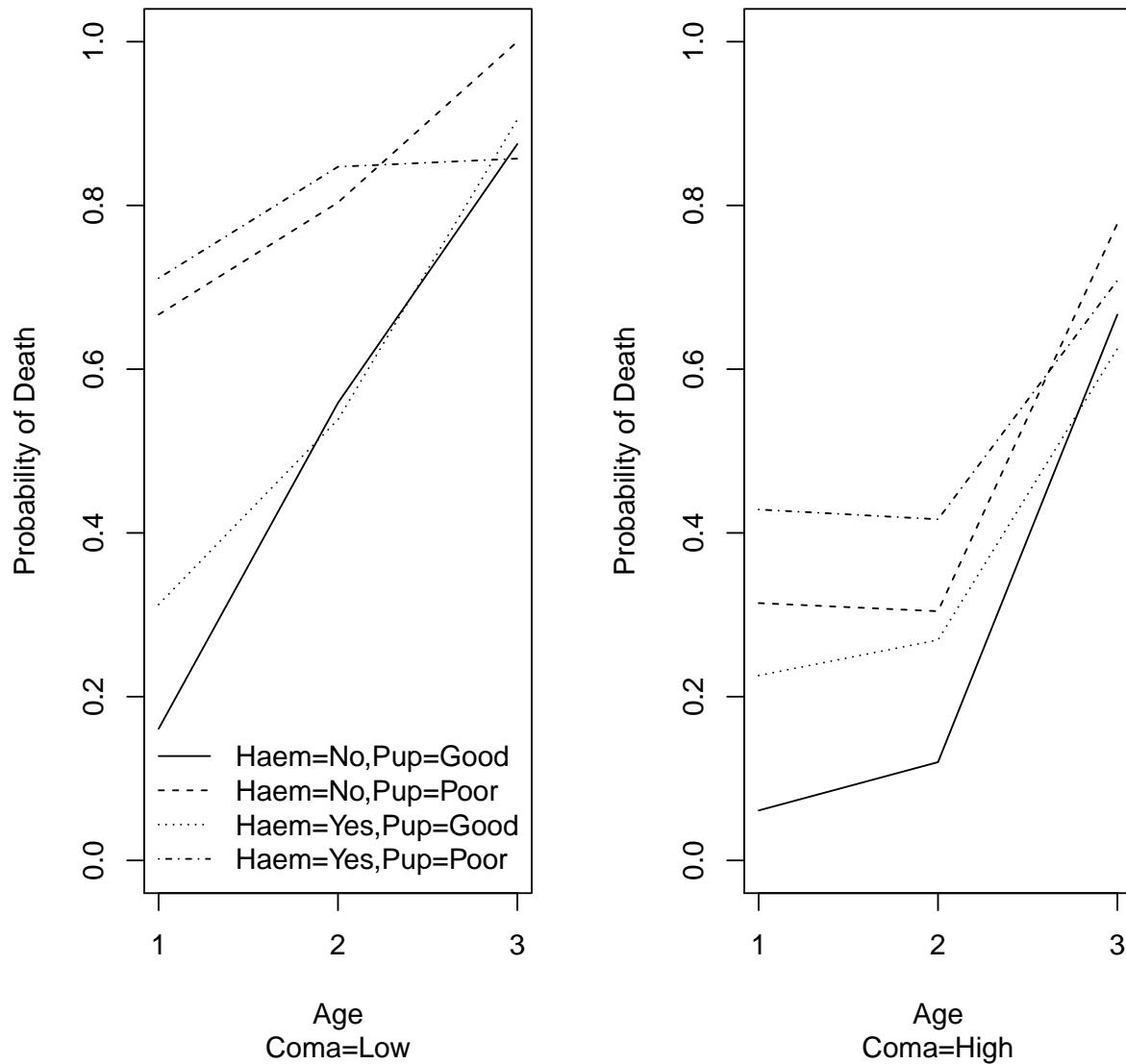
These data are counts ( $[y_i]$ ,  $i = 1, \dots, n$ ) of survival outcomes after head injury (dead/alive, 0/1) out of  $n_i$  outcomes for patients categorized into category  $i = 1, \dots, n = 24$  (encoded in covariate vector  $[\mathbf{x}_i]$ ) shortly after injury.

See [Wak13, Tab. 1.1, p. 9] or the following output.

| Age   | Coma | Hematoma | Pupils | Outcome |       |
|-------|------|----------|--------|---------|-------|
|       |      |          |        | Dead    | Alive |
| 1–25  | Low  | No       | Good   | 9       | 47    |
|       |      |          | Poor   | 58      | 29    |
|       | Yes  | No       | Good   | 5       | 11    |
|       |      |          | Poor   | 32      | 13    |
|       | High | No       | Good   | 5       | 77    |
|       |      |          | Poor   | 11      | 24    |
|       |      | Yes      | Good   | 7       | 24    |
|       |      |          | Poor   | 12      | 16    |
| 26–54 | Low  | No       | Good   | 19      | 15    |
|       |      |          | Poor   | 45      | 11    |

|      |     |      |      |    |    |
|------|-----|------|------|----|----|
|      |     | Yes  | Good | 21 | 18 |
|      |     |      | Poor | 61 | 11 |
| High | No  | Yes  | Good | 6  | 44 |
|      |     |      | Poor | 7  | 16 |
| >=55 | Low | Yes  | Good | 14 | 38 |
|      |     |      | Poor | 15 | 21 |
|      |     | No   | Good | 7  | 1  |
|      |     |      | Poor | 20 | 0  |
|      |     | Yes  | Good | 19 | 2  |
|      |     |      | Poor | 42 | 7  |
|      |     | High | Good | 12 | 6  |
|      |     |      | Poor | 7  | 2  |
|      |     | Yes  | Good | 25 | 15 |
|      |     |      | Poor | 17 | 7  |

Some plots, below, corresponding to the table, above.



A possible model for these data:

$$y_i \mid p_i \sim \text{binomial}(n_i, p_i),$$

as arising from  $y_i = \sum_{j=1}^{n_i} z_{ij}$  where

$$z_{ij} \sim \text{Bernoulli}(p_i)$$

(though see [Wak13, Chap. 7, p. 309]),  $p_i = P(Z_{ij} = 1 \mid \mathbf{x}_i)$ , and

$$\text{logit}(p_i) = \mathbf{x}_i^t \boldsymbol{\beta}_i,$$

likely including an intercept term and interaction terms beyond main effects (see, e.g., [Wak13, pp. 322-3]), and we take  $p_i$  to be the probability of survival, so that  $z_{ij}$  indicates survival (1) for the  $j$ th patient in the  $i$ th category and  $y_i$  counts surviving patients in category  $i$ .

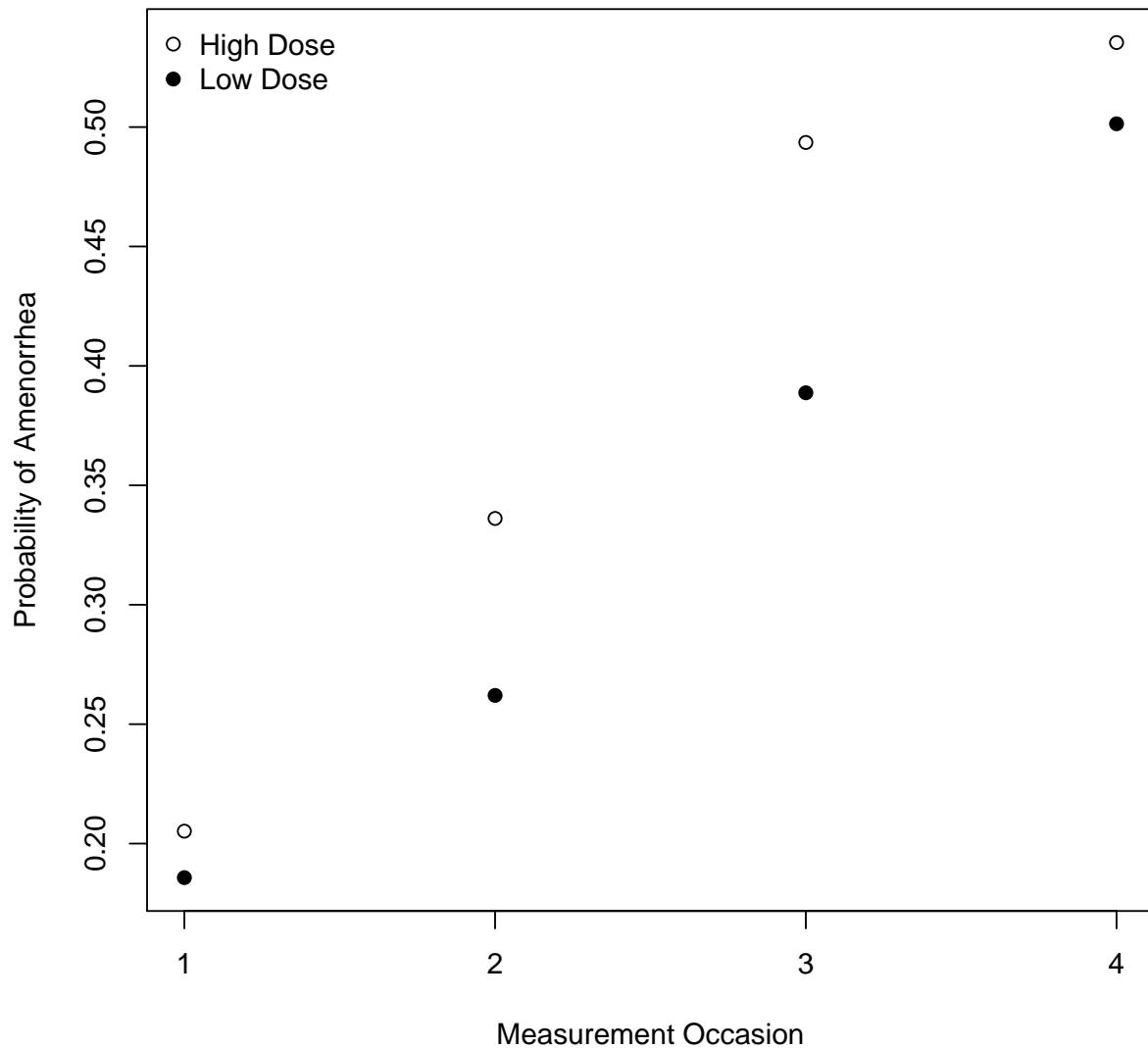
### Some Considerations for These Head Injury Data:

1. Scope of inference?
2. Chapter 7 Binary Data models (independent data). Variable selection.
3. Part IV Nonparametric (Predictive) Modeling (Chapters 10-12).

## 2.4 Contraception Drug

These data consist of the binary (0/1) response ( $[y_{ij}]$ ) over time ( $[t_{ij}]$ ), indicating amenorrhea (absence of menstrual bleeding) for observation  $j$  of woman  $i$  assigned to one of two dose groups ( $[d_i]$  0/1, low/high) in a randomized longitudinal trial involving the contraception drug DMPA.

The figure below shows empirical average responses (proportions, right!?) for the women at each combination of dose and time. (See [Wak13, Fig. 9.1, p. 427].)



Below are the empirical correlations and variances of responses for each group of women over time (low dose then high dose). (See [Wak13, Tab. 9.1, p.427].)

|      | [,1]      | [,2]      | [,3]      | [,4]      |
|------|-----------|-----------|-----------|-----------|
| [1,] | 1.0000000 | 0.4003459 | 0.2832044 | 0.2743381 |
| [2,] | 0.4003459 | 1.0000000 | 0.4501704 | 0.3543551 |
| [3,] | 0.2832044 | 0.4501704 | 1.0000000 | 0.5253738 |

```
[4,] 0.2743381 0.3543551 0.5253738 1.0000000
[1] 0.1515187 0.1937882 0.2382065 0.2506925
      [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.3078897 0.2523948 0.2852843
[2,] 0.3078897 1.0000000 0.4335164 0.4321697
[3,] 0.2523948 0.4335164 1.0000000 0.4720421
[4,] 0.2852843 0.4321697 0.4720421 1.0000000
[1] 0.1633874 0.2236179 0.2506029 0.2494527
```

A model for these data follows. (See [Wak13, Sec. 9.13.3, p. 462].)

$$y_{ij} \sim_{ind} \text{Bernoulli}(p_{ij}),$$

with

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = (\beta_0 + b_i) + (\beta_1 + \beta_2 d_i)t_{ij} + (\beta_3 + \beta_4 d_i)t_{ij}^2.$$

### Some Considerations for these Contraception Drug Data:

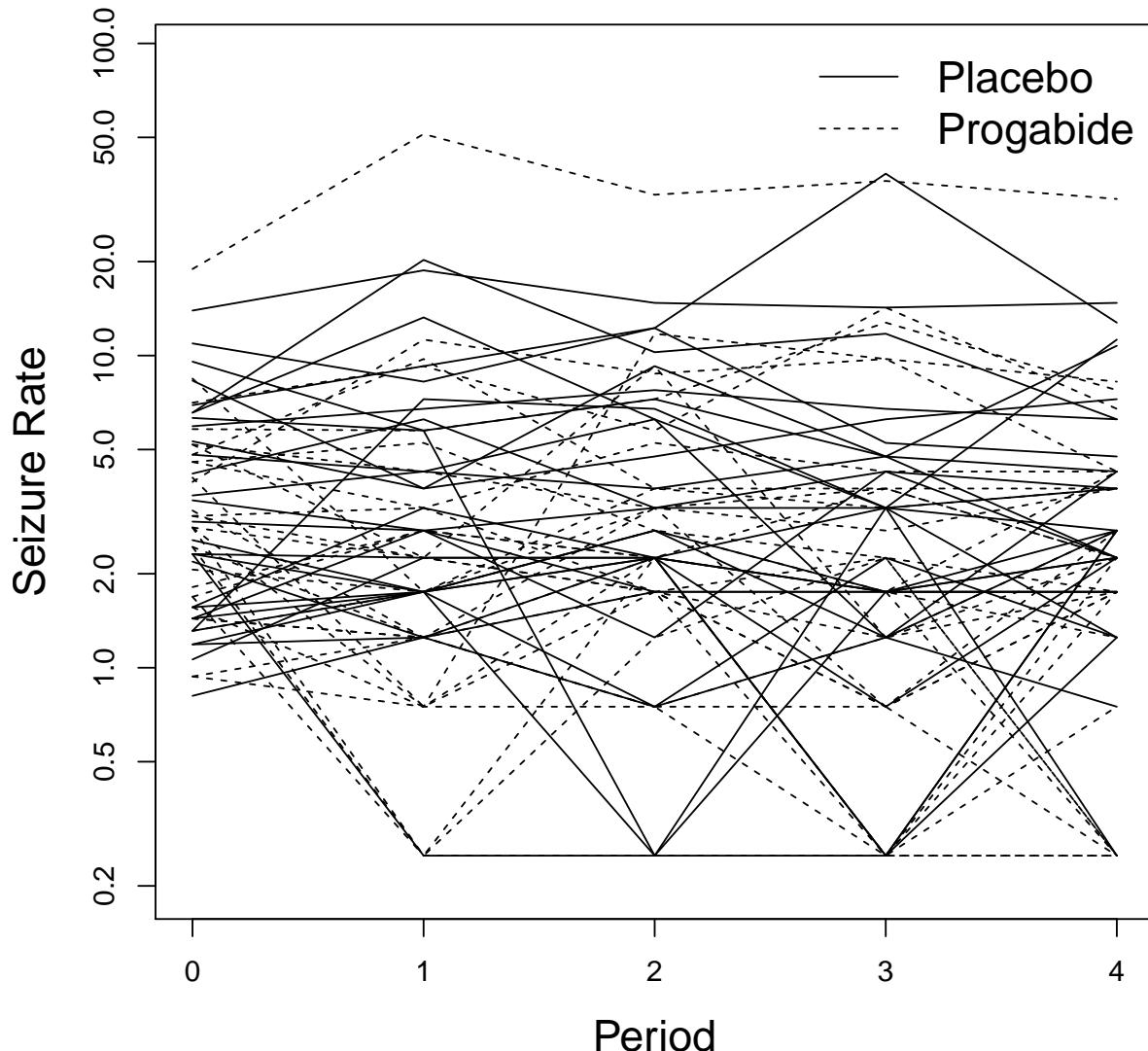
1. Scope of inference?
2. How are the correlations and variances above computed?
3. Why might we not have different intercept effects for the different groups?
4. Multivariate binary data (correlation over time within woman).
5. Chapter 9 General Regression Models (dependent data). NLMM. GEE.

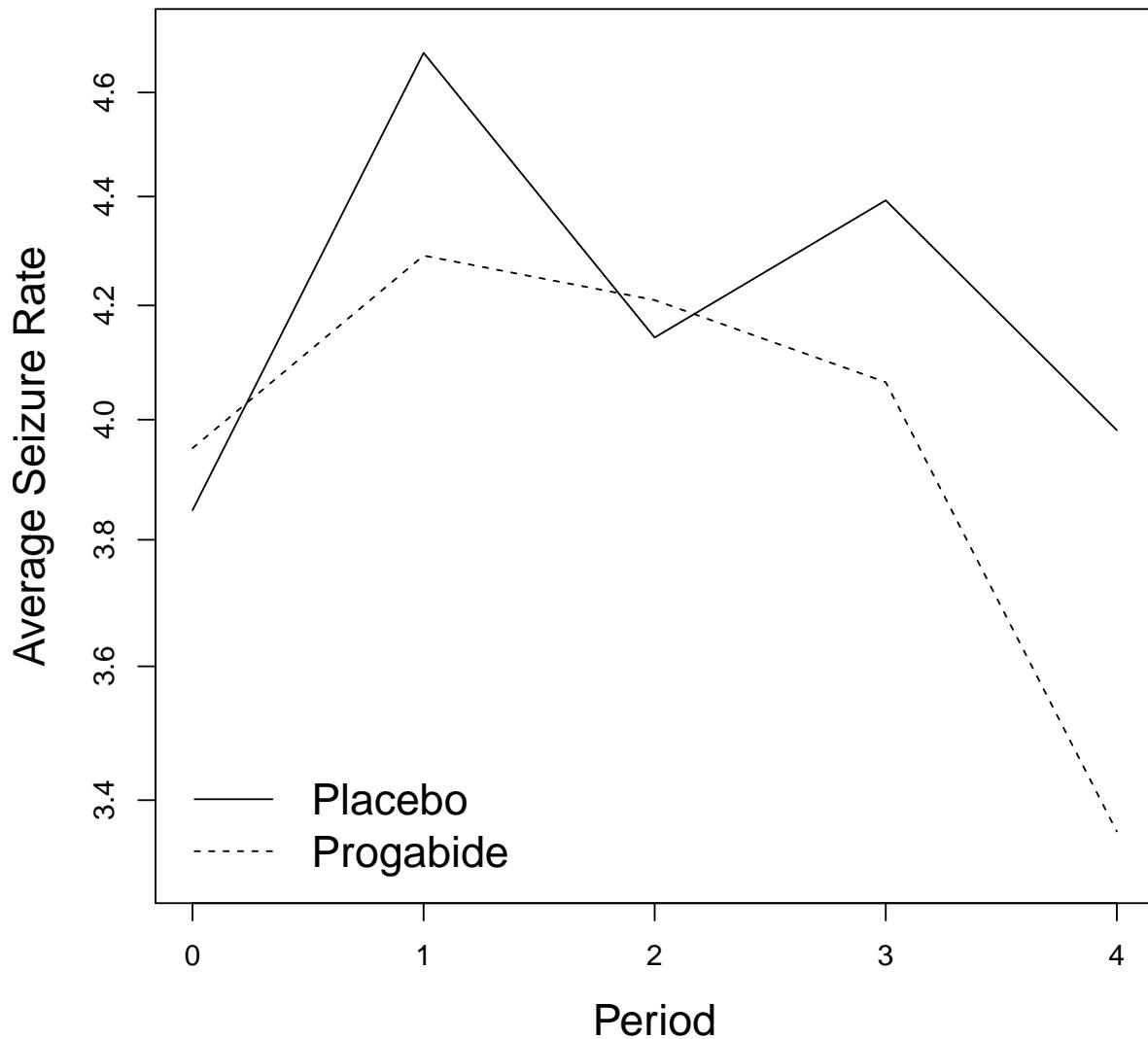
## 2.5 Seizure Data

These data consist of the observed count ( $y_{ij}$ ) of seizures during the  $j$ th two-week observation period ( $T_{ij}$ ) ( $T_{i0} = 2$ ,  $T_{ij'} = 8$  for  $j' = 1, \dots, 4$ ) for

individual  $i$  assigned to one of two drug groups ( $x_{1i}$  0/1, placebo/drug) in a randomized longitudinal trial involving the antiepileptic drug progabide. Also,  $x_{2j}$  indicates (0/1, pre/post) post-treatment periods ( $T_{i0} = 2$  is the pre-treatment period and  $T_{ij'} = 8$  are post-treatment), regardless of treatment group, and  $x_{3ij} = x_{1i}x_{2j}$  indicates pre/post for the active drug group (1=if progabide group at post treatment ( $x_{1i} = 1$  and  $x_{2j} = 1$ ), 0 otherwise ( $x_{1i} = 0$  or  $x_{2j} = 0$ )).

The following plots show the empirical seizure rates ( $y_{ij}/T_{ij}$ ) and their averages by combination of period number and drug group. (See [Wak13, Figs. 9.2 & 9.3, pp. 428-9].)





The following shows the ratios of empirical variance of seizure count to the empirical average, by treatment group and period number  $j$ , placebo group results followed by the progabide group results. (See [Wak13, Tab. 9.2, p. 429].

```
[1] 22.1 11.0  8.0 24.5  7.3  
[1] 24.8 38.8 16.7 23.7 18.9
```

A model for these data may be

$$y_{ij} \sim_{ind} \text{Poisson}(\mu_{ij}),$$

where

$$\mu_{ij} = T_{ij} \exp(\mathbf{x}_{ij}^t \boldsymbol{\beta} + b_{i0}),$$

which we may write as

$$\log(\mu_{ij}) = \log(T_{ij}) + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2j} + \beta_3 x_{ij} + b_{i0},$$

where  $\mathbf{x}_{ij}$  collects the above mentioned covariates and a “1” into a vector. Incidentally,  $T_{ij}$  (or  $\log(T_{ij})$ ) is what’s often called and “offset,” a known quantity in a linear predictor. (See, [Wak13, pp. 427-8 & pp. 433-4].)

### Some Considerations for these Seizure Data:

1. Scope of inference?
2. Dependence of counts within the same individual.
3. Overdispersion (extra-Poisson Variation).
4. Chapter 9 General Regression Models (dependent data). GLMM Poisson model. Extra-Poisson variation (overdispersion). Dependence within individual. Temporal dependence?

## 2.6 Lung Cancer & Radon

These data investigate the relationship between standardized morbidity ratio ( $\boxed{SMR_i}$ ), for lung cancer during the period 1998-2002, with average residential radon concentration ( $\boxed{x_i}$ ,  $pCi/l$ ) for counties in Minnesota.

$$\boxed{SMR_i = \frac{y_i}{E_i}},$$

where  $y_i$  is the lung cancer incidence (counts of males and females in several age groups) in county  $i$  and  $E_i$  is an “expected” count estimated as

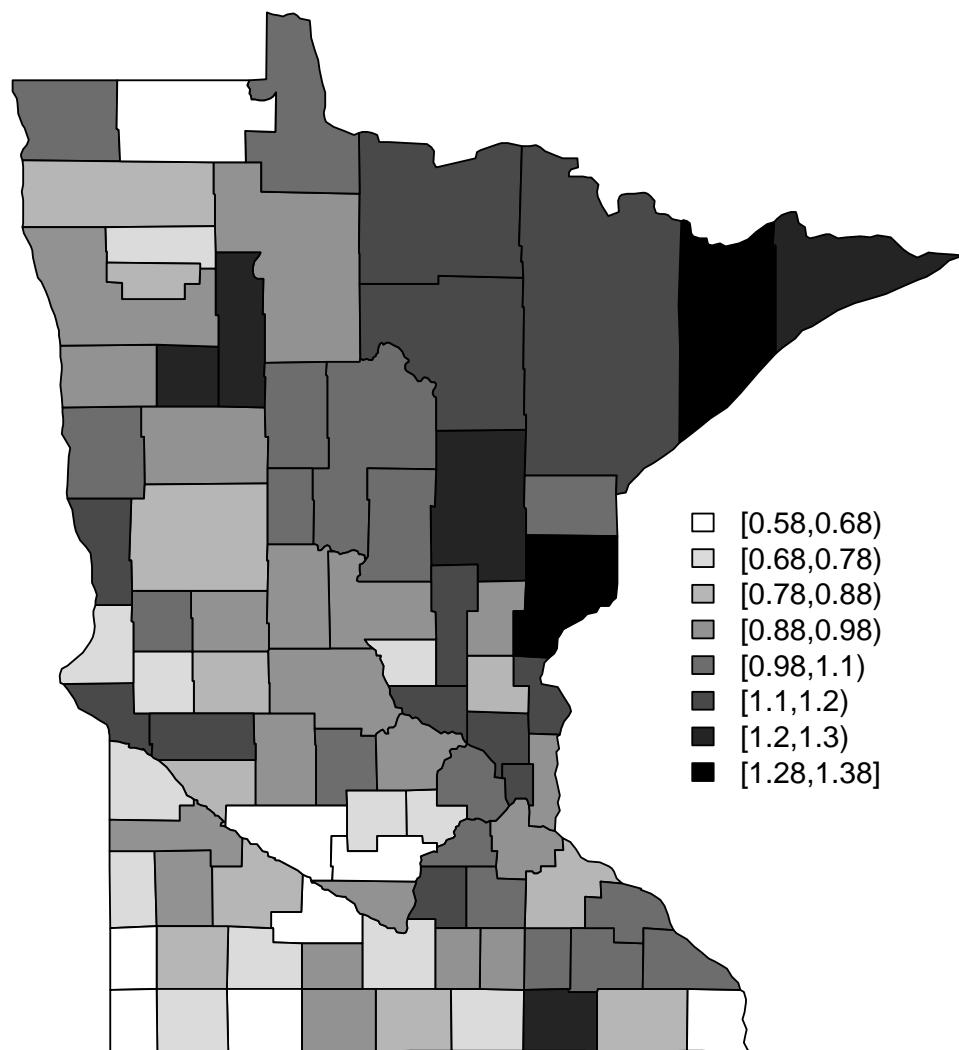
$$E_i = \sum_{j=1}^J N_{ij} q_j,$$

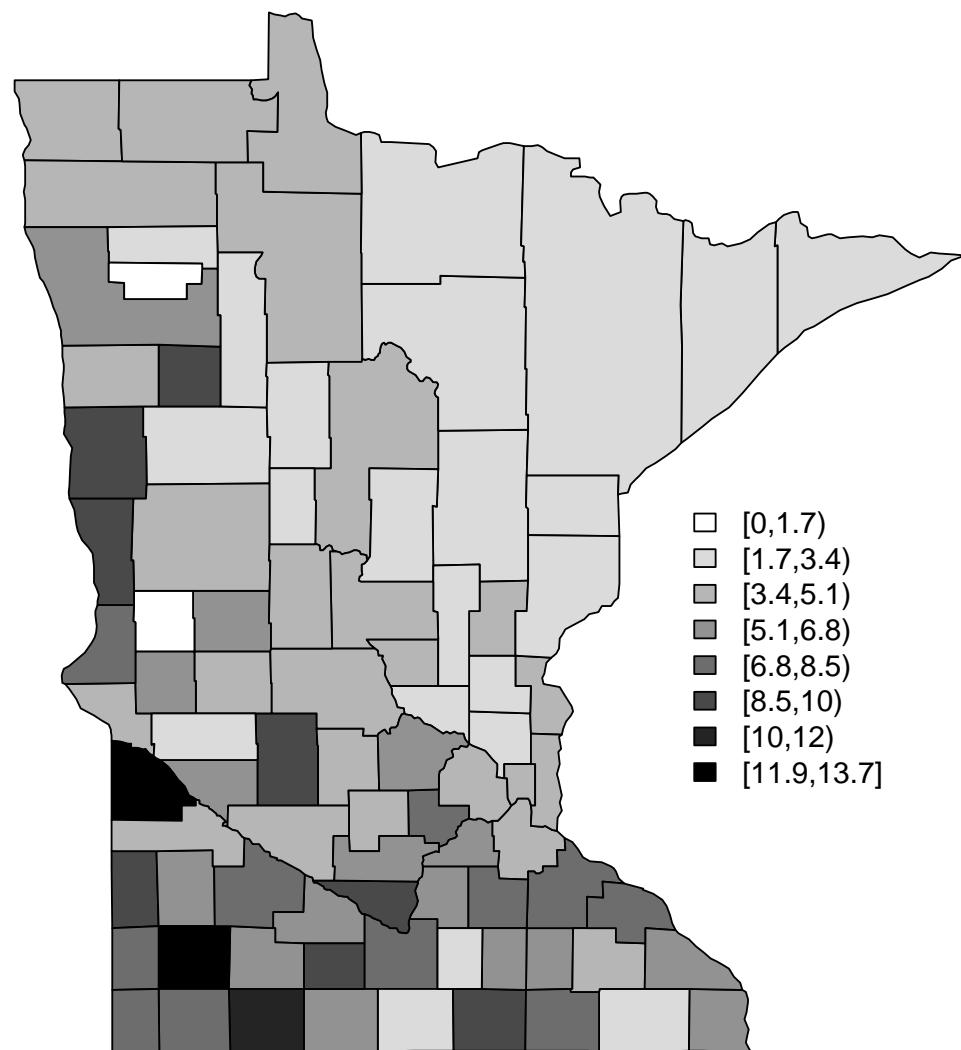
and  $N_{ij}$  is the population in stratum  $j$  (gender and age categories) in county  $i$ , and  $q_j$  is a “reference” probability of lung cancer for stratum  $j$  (gender and age categories).

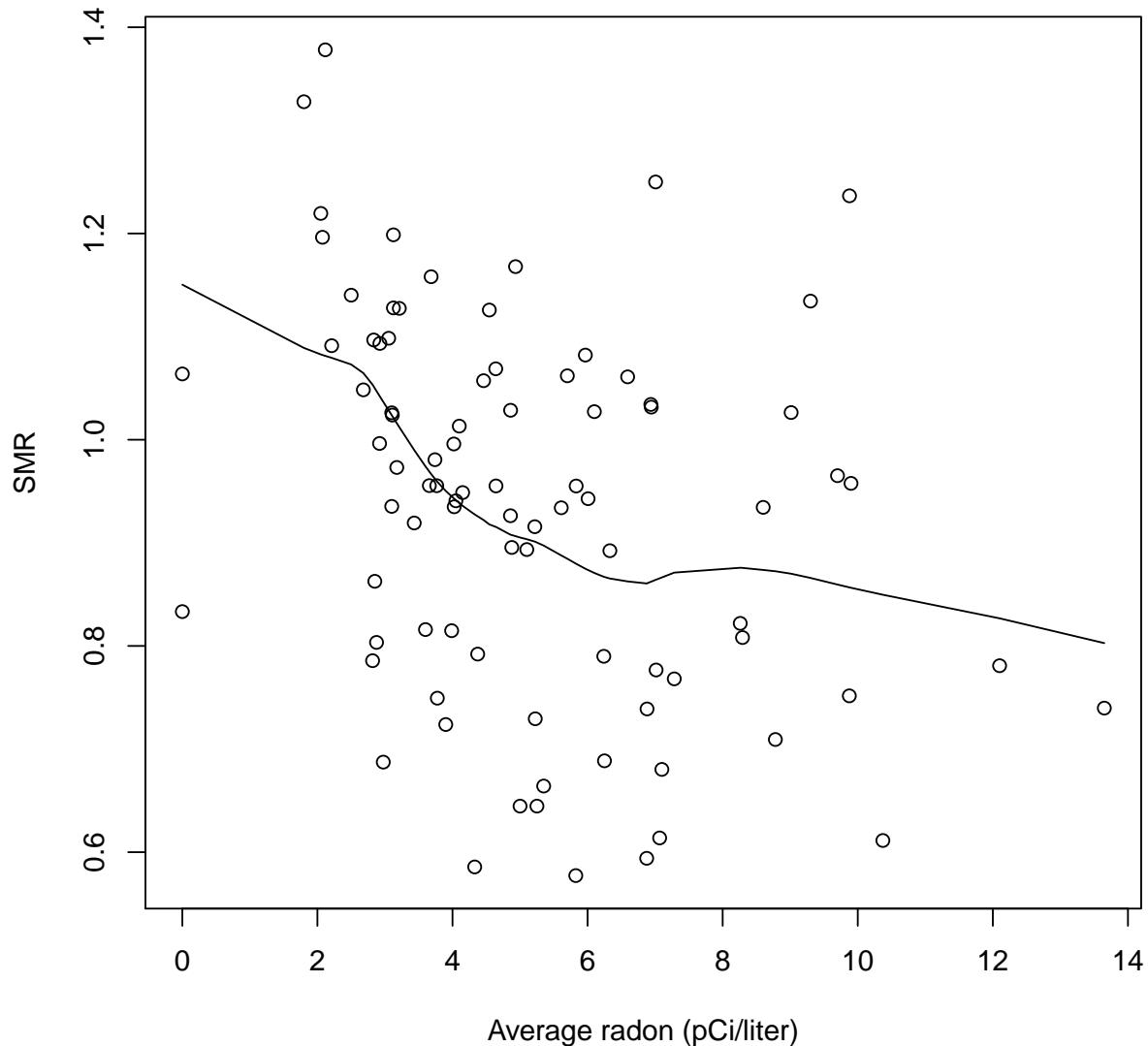
Thus,  $E_i$  is an “expected count” of lung cancer cases in county  $i$ , and we should see  $SMR_i$  as indicating higher/lower counts than expected as  $SMR_i > 1$  or  $SMR_i < 1$ , respectively (controlling for population size, importantly!). (This is suggested, roughly, by  $\text{binomial}(N_{ij}, q_j)$ .)  $E_i$  (or  $\log(E_i)$ ) is an example of what is referred to as an “offset,” a known quantity in a statistical model. More later of course.

(NOTE:  $N_{ij}q_j$  appears to be precomputed in the data set for males and females, these two expected counts then summed to get  $E_i$  here; again, see code/data via author’s web site.)

Plots follow: SMR map, radon map, then the scatterplot of SMR vs. radon.







A model considered in [Wak13, Sec. 1.3.4] is

$$y_i \sim_{ind} \text{Poisson}(\lambda_i), \quad \text{where}$$

$$\log\left(\frac{\lambda_i}{E_i}\right) = \log E\left[\frac{y_i}{E_i} \mid x_i\right] = \beta_0 + \beta_1 x_i,$$

a Poisson GLM model with log link. Notice the aforementioned known “offset” term.

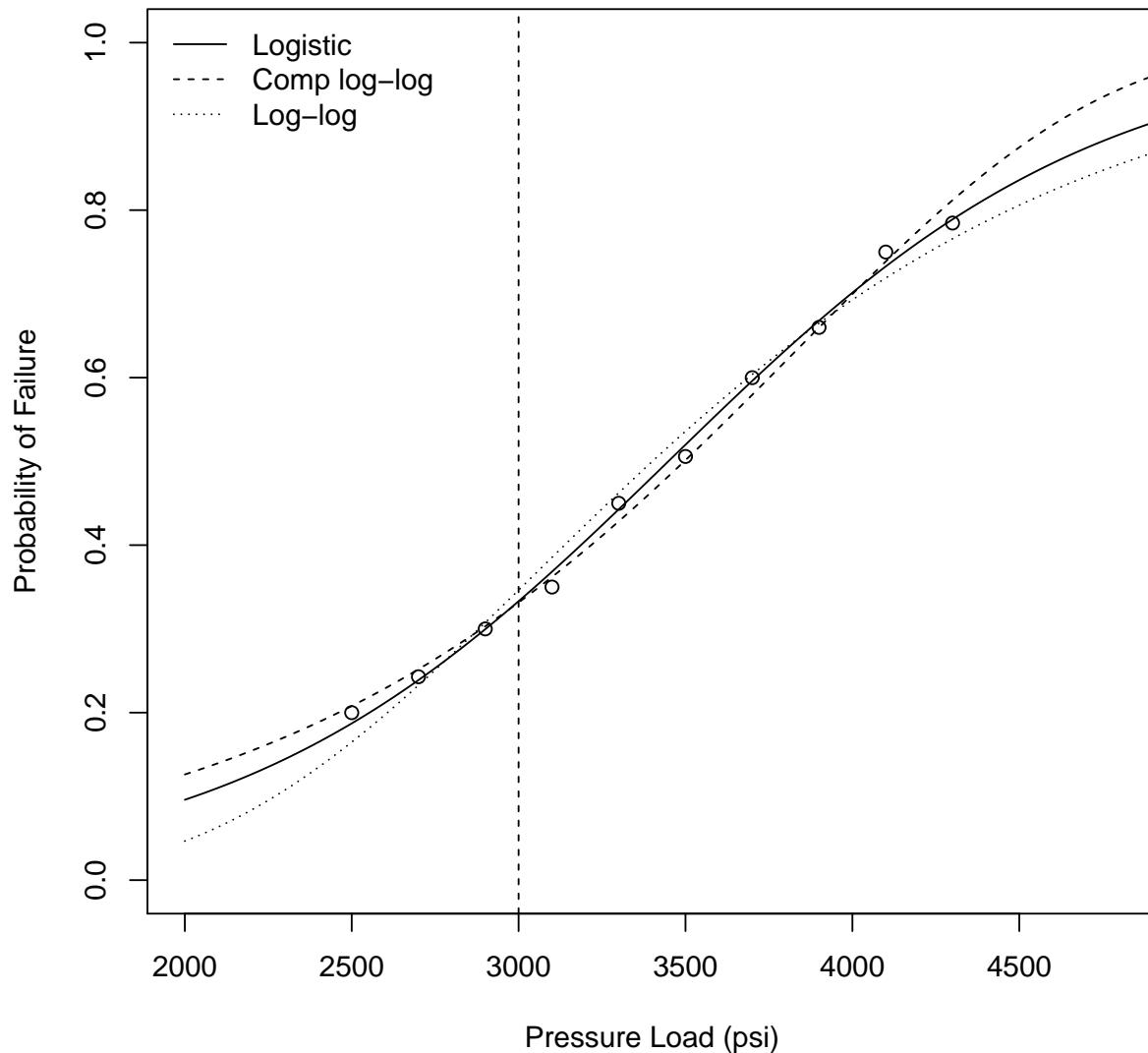
**Some Considerations for these Lung Cancer Data:**

1. Scope of inference?
2. Overdispersion (or “excess Poisson variation” in this case). Why might we expect this? → Poisson-gamma model or negative binomial, later.
3. Ecological (aggregation) bias or fallacy (Figure 1.5).
4. Smoking status?
5. Chapter 6 General Regression Models (independent data)
6. Chapter 9 General Regression Models (dependent data) (spatial dependence here).

## 2.7 Aircraft Fastener Data

These data are from a study of the compressive strength of aircraft fasteners wherein  $\boxed{y_i}$  is the number of fasteners that failed out of  $\boxed{n_i}$  fasteners subjected to compressive pressure  $\boxed{x_i}$  (psi),  $i = 1, \dots, n$ .

The following plot of the data and some model fits (discussed in class and below) essentially reproduces [Wak13, Fig. 7.4, p. 325].



A model for these data is

$$y_i \mid p_i \sim_{ind} \text{binomial}(n_i, p_i) \quad \text{where}$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i.$$

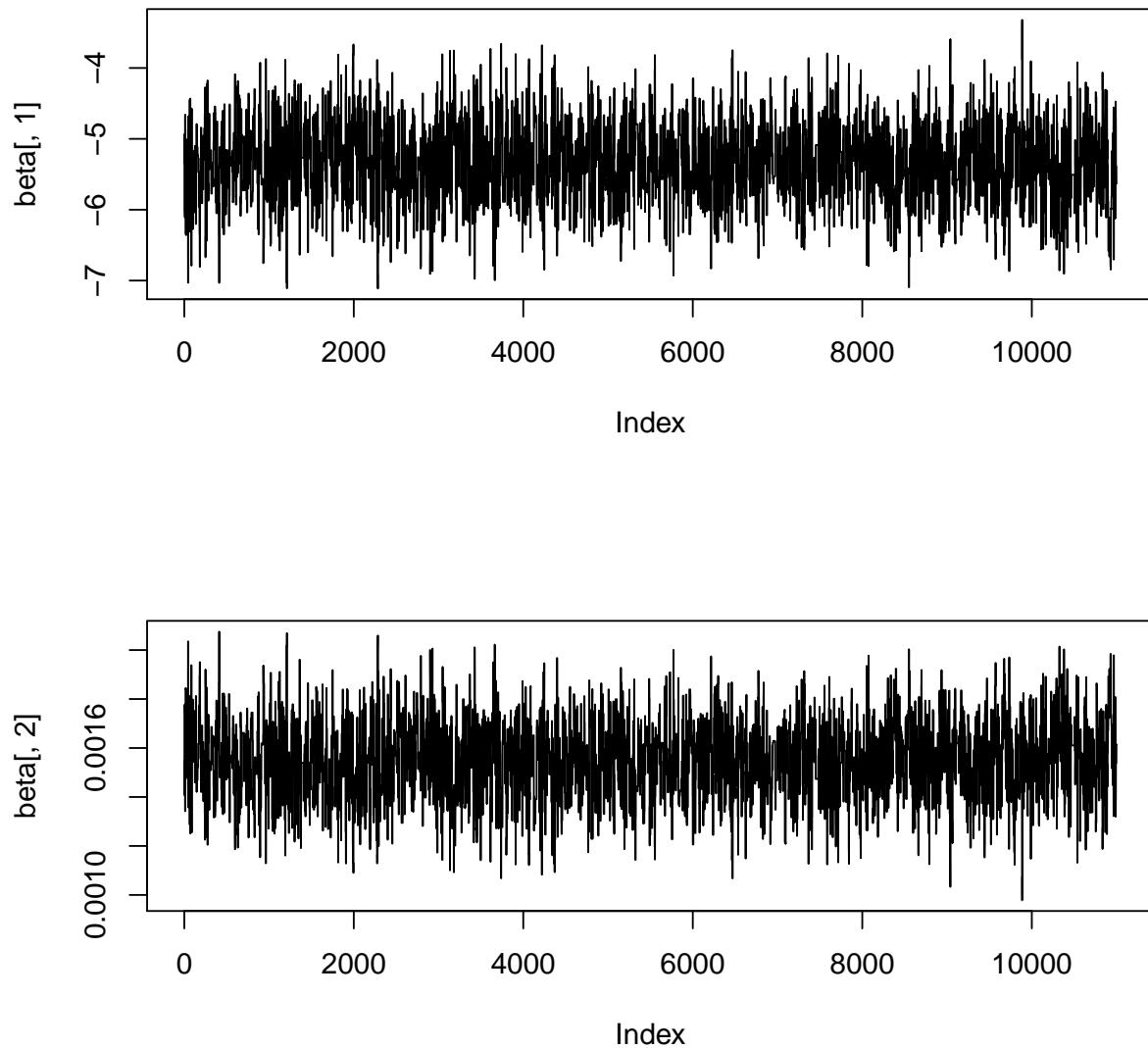
See [Wak13, pp. 323+]. (This model was fitted using MLE—the solid line in the above plot, after taking the expit transform of the linear predictor to get to the  $p$  scale.)

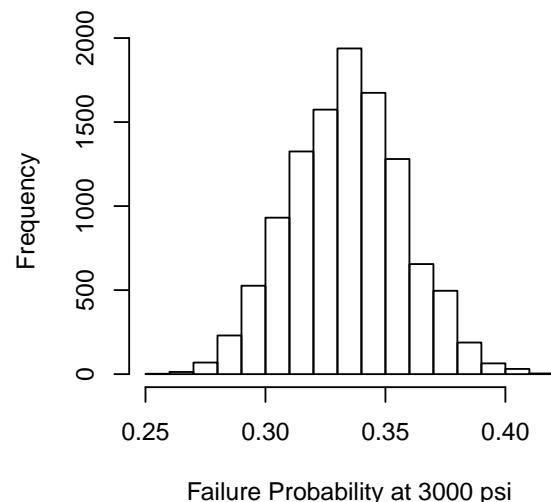
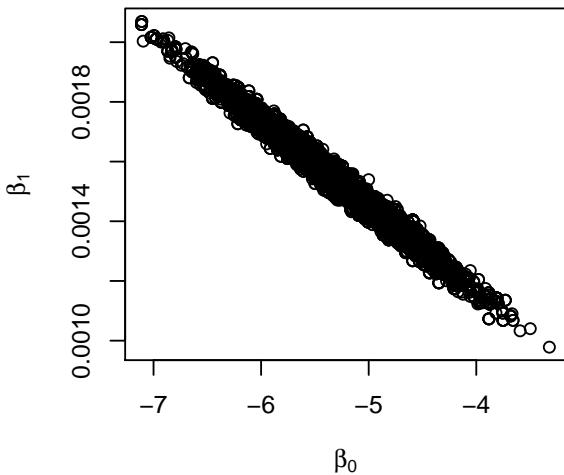
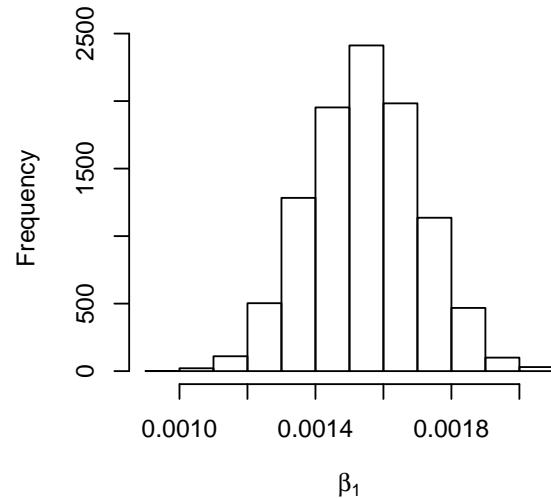
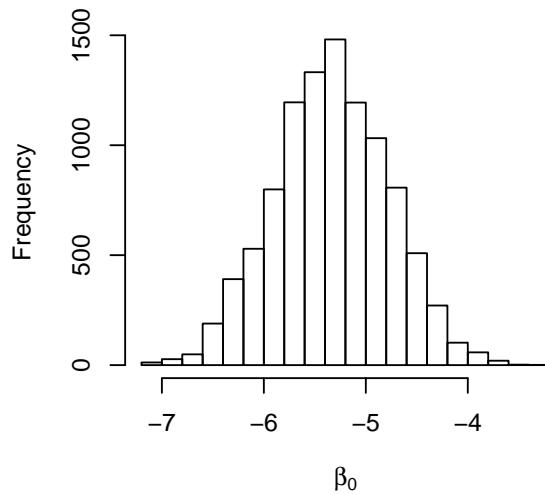
For comparison, the next chunk shows a short summary of MCMC computations for a Bayesian analysis of the data with the above (logit link) model.

```
Accept = 0.297
beta0 2.5 50 97.5: -6.424113 -5.346292 -4.271304
beta1 2.5 50 97.5: 0.00123583 0.001549736 0.001855061
exp(beta1) 2.5 50 97.5: 1.001237 1.001551 1.001857
```

Using results obtained in the previous chunk, the next chunk summarizes the posterior distribution graphically, including, at the very end, a 95% credible interval for the probability of failure at  $x = 3000$  (psi), i.e., for

$$p(x = 3000) = \frac{\exp(\beta_0 + \beta_1 3000)}{1 + \exp(\beta_0 + \beta_1 3000)}.$$





95% CI for  $p(3000)$ : 0.2886972 0.3801869

### Some Considerations for these Fastener Data:

1. Scope of inference?

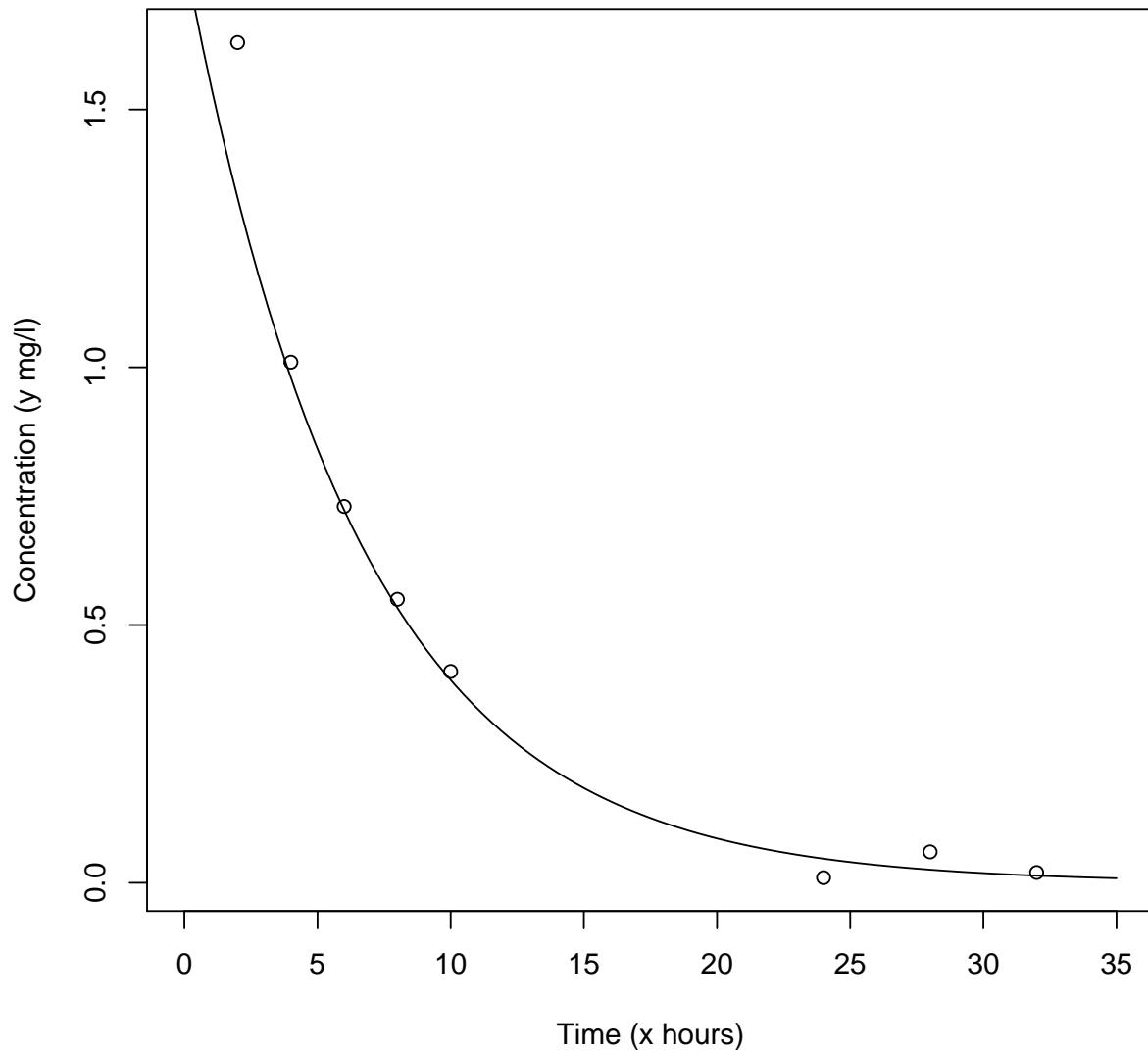
2. These data are also used to quickly illustrate the use of Pearson residuals in this GLM context for comparing three link functions (associated fitted curves shown in a plot, above)
3. Chapter 7 Binary Data Models

## 2.8 Cardiac Failure & Cadralazine Concentration

These data consist of the  $i$ th measured concentration ( $y_i$ , mg/l) of the drug, cadralazine, in the blood of a cardiac failure patient at time  $x_i$  hours after administration of a  $d = 30$  mg dose of the drug, presumably by injection (not oral).

The following chunk shows a plot of the data, with a fitted curve (see next chunk)

```
> # Data from Table 6.5 for exercise 6.3.
> pkcadral.df<- data.frame(Time = c(2,4,6,8,10,24,28,32),
+                               Conc = c(1.63,1.01,0.73,0.55,
+                                       0.41,0.01,0.06,0.02))
> plot(Conc ~ Time, data = pkcadral.df,
+       xlim=c(0,35), xlab="Time (x hours)",
+       ylab="Concentration (y mg/l)")
>
> ## Add fitted curve (from previous hidden chunk and next chunk)
> curve(30 / pk.coef[1] * exp(- pk.coef[2] * x),
+        from=0, to=35, add=TRUE)
```



A simple, nonlinear model for such pharmacokinetic data is

$$\begin{aligned}\log y_i &= \mu_i(\boldsymbol{\beta}) + \epsilon_i \\ &= \log \left[ \frac{d}{v} \exp(-k_e x_i) \right] + \epsilon_i\end{aligned}$$

which we fit quickly in the following chunk:

```
> # Data from Table 6.5 for exercise 6.3.  
> pkcadral.df<- data.frame(Time = c(2,4,6,8,10,24,28,32),  
+                               Conc = c(1.63,1.01,0.73,0.55,  
+                                     0.41,0.01,0.06,0.02))  
> pkcadral.form<- log(Conc) ~ log(30 / v * exp(-ke * Time))  
> ## Crude self-starting value:  
> pkcadral.nls<- nls(pkcadral.form, data=pkcadral.df)  
  
Warning in nls(pkcadral.form, data = pkcadral.df): No starting values  
specified for some parameters.  
Initializing 'v', 'ke' to '1.'.  
Consider specifying 'start' or using a selfStart model  
  
> (pk.coef<- coef(pkcadral.nls))  
  
      v          ke  
16.6633093  0.1521064
```

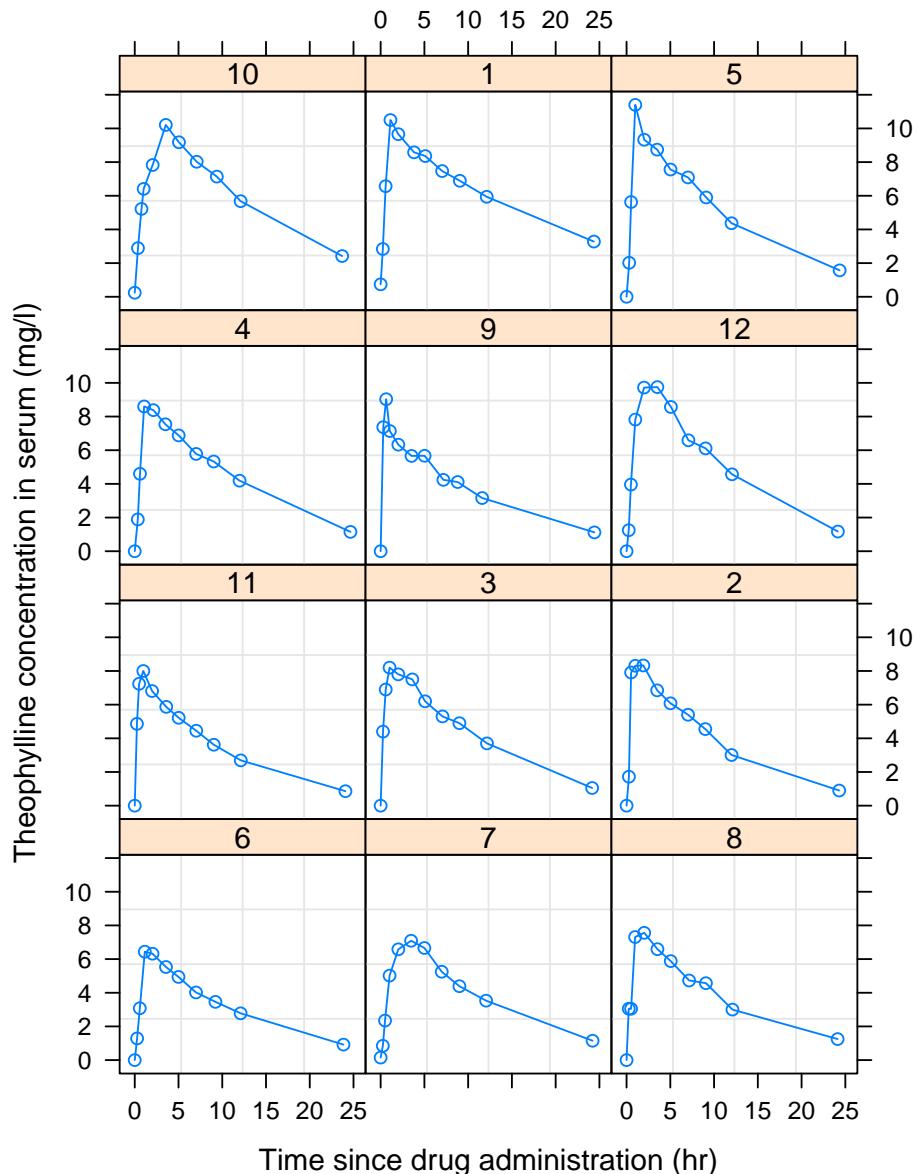
### Some Considerations for these Drug Concentration Data:

1. Scope of inference?
2. Nonlinear model for independent data ([Wak13, Chap. 6]).
3. Residual correlation over time? Nonlinear models for dependent data ([Wak13, Chap. 9]).
4. Similar to the Theophylline data, below, but those data arose from oral dosing, not injection, hence their different model using two compartments 0 (digestive system) and 1 (blood). (And, we consider only one individual here.)
5. See [Wak13, Exer. 6.3, p. 300] for these data.

## 2.9 Pharmacokinetics of Theophylline Data

These data consist of the  $j$ th measurement of the concentration  $[y_{ij}]$  ( $mg/l$ ) of the drug theophylline at time  $[x_{ij}]$  (hours) after initial dose ( $[d_i]$  ( $mg/kg$ )) for subject  $i$ .

Some plots follow.



The following essentially reproduces the data for subject 1 given in [Wak13, Tab. 1.2, p. 13].

Grouped Data: conc ~ Time | Subject

|    | Subject | Wt   | Dose | Time  | conc  |
|----|---------|------|------|-------|-------|
| 1  | 1       | 79.6 | 4.02 | 0.00  | 0.74  |
| 2  | 1       | 79.6 | 4.02 | 0.25  | 2.84  |
| 3  | 1       | 79.6 | 4.02 | 0.57  | 6.57  |
| 4  | 1       | 79.6 | 4.02 | 1.12  | 10.50 |
| 5  | 1       | 79.6 | 4.02 | 2.02  | 9.66  |
| 6  | 1       | 79.6 | 4.02 | 3.82  | 8.58  |
| 7  | 1       | 79.6 | 4.02 | 5.10  | 8.36  |
| 8  | 1       | 79.6 | 4.02 | 7.03  | 7.47  |
| 9  | 1       | 79.6 | 4.02 | 9.05  | 6.89  |
| 10 | 1       | 79.6 | 4.02 | 12.12 | 5.94  |
| 11 | 1       | 79.6 | 4.02 | 24.37 | 3.28  |

A model arises from consideration of the pharmacokinetics of the drug in a body (see [Wak13, Fig. 1.7, p. 14]). Let  $w_k(x)$  be the *amount* of drug in body compartment  $k = 0, 1$ , where we may consider compartment 0 to be the digestive system (the drug is administered orally) and compartment 1 to be the blood. The drug flow (pharmacokinetics) may be modeled by

$$\begin{aligned}\frac{dw_0(t)}{dt} &= -k_a w_0(t) \\ \frac{dw_1(t)}{dt} &= k_a w_0(t) - k_e w_1(t),\end{aligned}$$

where parameter  $k_a > 0$  is the absorption rate of the drug into compartment 1 from compartment 0, and parameter  $k_e > 0$  is the elimination rate from compartment 1. This system of ODEs may be solved for  $w_1(t)$  by specifying an initial condition,  $w_0(0) = d$ , the initial dose amount of the drug (switching notation to time “x” now):

$$w_1(x) = \frac{dk_a}{k_a - k_e} [\exp(-k_e x) - \exp(-k_a x)].$$

Dividing by blood volume,  $v$  (another unknown parameter), gives the drug concentration in compartment 1 at time  $x$  (switching to usual statistical

notation):

$$\mu(x) = \frac{dk_a}{v(k_a - k_e)} [\exp(-k_ex) - \exp(-k_ax)].$$

We might complete our model specification with

$$y_{ij} = \mu(x_{ij}) + \epsilon_{ij},$$

and  $\epsilon_{ij} \sim_{iid} N(0, \sigma_\epsilon^2)$ .

### Some Considerations for these Theophylline Data

1. The response must be positive.
2. Nonconstant variance?
3. The above two considerations may *may* be mitigated by modeling

$$\log(y_{ij}) = \log(\mu(x_{ij})) + \delta_{ij},$$

etc.

4. Process model error (or “discrepancy”)? We might expect process model error to result in smooth departures from the actual pharmacokinetic dynamics, i.e., for the curve  $\mu(x)$  do depart smoothly from the actual concentration curve, and this sort of smooth departure may be modeled by a temporal correlation model (within each individual) (intuitively, to account for or “soak up” unmeasured covariation that may change smoothly over time).
5. Individual (random) curves?  $(v_i, k_{ai}, k_{ei})$ ?
6. Nonidentifiability of parameters. What is  $\mu(x)$  when we plug in the parameter value  $(vk_e/k_a, k_e, k_a)$  for  $(v, k_a, k_e)$  in the model for  $\mu(x)$ , above? The added constraint  $k_a > k_e > 0$  will take care of this. For example,  $k_a = k_e + a$ ,  $k_e > 0$  and  $a > 0$ , where, perhaps,  $a = \exp(lna)$ ,  $-\infty < lna < \infty$ . Or,  $k_a = ak_e$  where  $a = \exp(lnb) + 1$  and  $-\infty < lnb < \infty$ . (And, perhaps,  $k_e = \exp(lnke)$   $-\infty < lnke < \infty$ .)

7. You can sort of see intuitively the nonidentifiability. To illustrate, assume a larger elimination rate than assimilation rate, i.e.,  $k_e > k_a$ . For a fixed volume  $v$ , we get some concentration curve,  $\mu(x)$ . In order to maintain the same concentration in with a larger volume,  $vk_e/k_a$ , increase the assimilation rate relative to the elimination rate, which is done here exactly by swapping their values to get the same concentration curve  $\mu(x)$ . Thus, if you adopt the constraint  $k_e > k_a$ , there can be no such swapping in your estimation routine, no identifiability problem.
8. See [Wak13, Chap. 6, p. 255 & Chap.9, pp. 477-8] for more discussion of this nonidentifiable “flip-flop” model.
9. Other quantities of interest (see [Wak13, p. 15-16]) are no problem for a Bayesian approach, but generally present additional considerations for a frequentist approach.
10. These data are used a lot throughout Chapter 6 General Regression Models (independent data), where we try largely unsatisfactory alternative mean models, and throughout Chapter 9 General Regression Models (dependent data), where we consider most if not all of the above items.
11. Similar to the cadralazine drug concentration data [Wak13, Exer. 6.5, p. 303], discussed above, but caldralazine is, evidently, injected directly into the blood, hence the different models for these data sets.

## 2.10 Course/Textbook Outline

- [Wak13, Chap. 1 Introduction and Motivating Examples]. We use this textbook chapter for its presentation of some of the motivating examples discussed above, and used some of its discussion of fundamental concepts in our previous lecture chapter (1). We may return to some of this material later, but not likely in a systematic fashion.

- [Wak13, Part I, Inferential Approaches]

This part, including textbook chapters 2-4, discusses foundational methods, including philosophical and computational methods, used throughout the remainder of the text. Frequentist inferential methods (chapter 2) begin with likelihoods (model-based inference), and proceed to more empirical methods requiring fewer model assumptions (so-called “robust” methods), including quasi-likelihood and estimating equations, along with sandwich variance estimation in estimating equations. Chapter 3 discusses Bayesian inference methods, which is also likelihood-based, but now including prior probability models (again, model-based inference). Chapter 4 discusses hypothesis testing from both frequentist and Bayesian perspectives.

We will attempt largely to avoid covering these materials systematically, but will tend to introduce them selectively in the context of our coverage of Part II and, time permitting, some of Part IV.

- [Wak13, Part II, Independent Data]

- [Wak13, Chap. 5 Linear Models]. If there is a **main textbook chapter** for this course, this is it. It’s relatively brief, and we will see more detail in our notes based on other sources.
- [Wak13, Chap. 6 General Regression Models] Generalized linear models (GLM) and general nonlinear models for independent data. If we haven’t already begun to discuss variance modeling at this point, in addition to the primary regression enterprise of mean modeling, we will begin to focus more on variance models.

**Theophylline Data.** Gamma GLM via MLE; via quasi-likelihood (overdispersion) with sandwich variance estimator; and via GLM Bayes. Nonlinear one-compartment model (discussed above) via Gaussian MLE and via estimating equations with sandwich variance estimation; and via Bayes.

**Lung Cancer and Radon Data.** Poisson GLM via MLE with log link and linear link; via quasi-likelihood (overdispersion); and via Bayes. (Could also have considered a Poisson-Gamma model

or negative binomial model in either and MLE or Bayes approach (which may be in earlier chapters...).

- [Wak13, Chap. 7 Binary Data Models] A specialize look at models for binary data, which may have easily been included in the GLM part of Chapter 6.

**Head Injury Data.** GLM via MLE and via Bayes, each with log link).

**Aircraft Fastener Data.** GLM via MLE using log link, cloglog link and loglog link; and via Bayes GLM with log link.

- [Wak13, Part III, Dependent Data] Dependence in our textbook typically refers to dependence among measurements *within* an “individual” unit (i.e., “subject”), and we almost always have independent units in this book/course (with exception perhaps of the lung cancer/radon data as treated in chapter 9). Part III material will be the primary focus of INF 512 Modern Regression II.

- [Wak13, Chap. 8 Linear Models] Linear mixed models (LMM).

**Dental Growth Data.** LMM via MLE/RMLE/Bayes. GEE for marginal model using independent and exchangeable working covariance models with model-based and sandwich covariance estimation.

**FEV1 Data.** LMM via MLE/REML/Bayes (AR(1) via MLE). GEE.

- [Wak13, Chap. 9 General Regression Models] Generalized linear mixed models (GLMM) and general nonlinear models, now for dependent data.

**Contraception Data.** GLMM via MLE (Laplace approximation and G-H quadrature as two different ways to marginalize out random effects) and via Bayes. GEE using independence and exchangeable working covariance models. And, alternating logistic regression (ALR).

**Seizure Data.** GLMM via MLE (Laplace approximation and G-H quadrature as two different ways to marginalize out random effects); via Bayes; and via conditional MLE. GEE with independent,

exchangeable, and AR(1) working covariance models (compared to Poisson and quasi-likelihood).

**Theophylline Data.** NLMM via MLE (normal); and via Bayes normal, lognormal, and power (?) first stage distributions. GEE.

**Lung Cancer and Radon Data.** GLMM via Bayes with spatial ICAR prior (INLA). (Comparison to (independent) Poisson, quasi-likelihood (overdispersion); negative binomial, GLMM non-spatial...estimation?...see code?)

- [Wak13, Part IV, Nonparametric Modeling]. This part (chapters 9-11) may also be called **semiparametric modeling** or **statistical learning** or **machine learning**, though essentially all of what we do in INF 511/512, at some level, has been usurped by “machine learning.” We hope to get to some of this material in both 511 and 512, which is almost entirely focused on independent data (like Part II).

### 2.10.1 Summary of This Course

Our textbook for this course brings together both frequentist and Bayesian methods in a relatively unique way, thanks mostly to the work of our text’s author, Jon Wakefield. Some of these methods are more or less “new,” relative to traditional (non)linear regression models, but it is the collection of these methods, together, in an effective and complimentary manner, that gives the course(s)/text its novelty. The text seeks to strike a balance between data analysis and theory, centered on methods and their comparisons.

# Lecture 3

## Basic Results in Probability and Statistics

### Contents

---

|  |            |
|--|------------|
| <b>3.1 Summations &amp; Products . . . . .</b>           | <b>83</b>  |
| 3.1.1 Summation Operator . . . . .                       | 83         |
| 3.1.2 Double Summation Operator . . . . .                | 85         |
| 3.1.3 Product Operator . . . . .                         | 85         |
| <b>3.2 Random Variables . . . . .</b>                    | <b>86</b>  |
| <b>3.3 Characteristics of Random Variables . . . . .</b> | <b>100</b> |
| 3.3.1 Expected (Mean) Value . . . . .                    | 100        |
| 3.3.2 Variance Operator . . . . .                        | 104        |
| <b>3.4 Random Vectors . . . . .</b>                      | <b>106</b> |
| 3.4.1 Covariance Operator & Its Properties . . . . .     | 106        |
| 3.4.2 Independence . . . . .                             | 109        |
| <b>3.5 Central Limit Theorem . . . . .</b>               | <b>110</b> |
| <b>3.6 Linear Functions of an RV . . . . .</b>           | <b>111</b> |
| <b>3.7 Linear Combinations of RVs . . . . .</b>          | <b>112</b> |

---

***Main Objectives:***

- Familiarize ourselves with basic results and notation to be used throughout the semester.

- Don't worry, these results are not the main focus of the course; we use them, not vice-versa.
- Learn R.

---

$\mathcal{O}$

***Additional Reading:***

Much of the material in this section of notes is based on [KNNL05, Appendix A]. We may change slightly some notation here. I have reserved a copy of [KNNL05] in the Cline Library for our use. You are welcome to read these references, of course, but, I suggest that our notes, here, may be sufficient. \_\_\_\_\_  $\mathcal{R}$

**NOTE:** We cover these basic results at the outset so that we can recall them later when needed during our discussion of regression and ANOVA, which is the real reason we are here, right?! Do your best to assimilate this information now, but do not become overly concerned about the apparent technical nature of the material. In particular, you will not be asked to derive the results despite some derivations shown here (i.e., you can just use the end results). You will not be asked to reproduce the results, by themselves, on an in-class exam, but you should know how to use them as they recur throughout the course in context of regression and ANOVA. We will use R to illustrate concepts here and throughout the course.

## 3.1 Summations & Products

### 3.1.1 Summation Operator

$$\sum_{i=1}^n Y_i = Y_1 + Y_2 + \cdots + Y_n$$

**NOTE:** We will often use uppercase letters toward the end of the alphabet to denote random variables (quantities that are uncertain before being observed; see definition, below) and their observed values. This is a bit different than many texts, which often try to distinguish random variables

from their observed values by using lowercase letters for the latter. In any case, summation is fundamental.

### Properties of Summation Operator

$$\sum_{i=1}^n k = \overbrace{k + k + \cdots + k}^{\text{n times}} = nk \quad k \text{ some constant}$$

$$\begin{aligned} \sum_{i=1}^n (Y_i + Z_i) &= Y_1 + Z_1 + \\ &\quad Y_2 + Z_2 + \\ &\quad \vdots \\ &\quad Y_n + Z_n + \\ &= \sum_{i=1}^n Y_i + \sum_{i=1}^n Z_i \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^n aY_i &= aY_1 + aY_2 + \cdots + aY_n \\ &= a(Y_1 + Y_2 + \cdots + Y_n) \\ &= a \sum_{i=1}^n Y_i \end{aligned}$$

Above results imply

$$\sum_{i=1}^n (a + bY_i) = na + b \sum_{i=1}^n Y_i$$

### 3.1.2 Double Summation Operator

$$\begin{aligned}
 \sum_{i=1}^n \sum_{j=1}^m Y_{ij} &= \sum_{i=1}^n (Y_{i1} + Y_{i2} + \cdots + Y_{im}) \\
 &= Y_{11} + Y_{12} + \cdots + Y_{1m} + \\
 &= Y_{21} + Y_{22} + \cdots + Y_{2m} + \\
 &= \vdots \\
 &= Y_{n1} + Y_{n2} + \cdots + Y_{nm} + \\
 &= \sum_{j=1}^m (Y_{1j} + Y_{2j} + \cdots + Y_{nj}) \\
 &= \sum_{j=1}^m \sum_{i=1}^n Y_{ij}
 \end{aligned}$$

### 3.1.3 Product Operator

$$\prod_{i=1}^n Y_i = Y_1 Y_2 \cdots Y_n$$

The following Chunk explores some of R's capabilities for computing sums and products.

```

> ### Concatenate numbers into an R vector:
> myvector<- c(5,6,9,2,9,4,8,2,1,4)
> (n<- length(myvector))

[1] 10

> myvector

[1] 5 6 9 2 9 4 8 2 1 4

```

```

> ### Name the elements because we care
> names(myvector) <- paste("y[", 1:n, "]", sep="")
> myvector

y[1]  y[2]  y[3]  y[4]  y[5]  y[6]  y[7]  y[8]  y[9]  y[10]
      5      6      9      2      9      4      8      2      1      4

> ### What's the 5th element myvector?
> ### "[" is an extraction operator.)
> myvector[5]

y[5]
9

> ### Sum and product of elements in myvector
> sum(myvector)

[1] 50

> prod(myvector)

[1] 1244160

```

## 3.2 Random Variables

- We gave a non-technical definition of a random variable in Definition 1.4.
- Intuitively, a random variable is a quantity that is uncertain before being observed (e.g., the weight of your cat, your blood serum cholesterol level, the next president of the United States, etc.).
- Yet the collection of possible values of a random variable behave in some structured manner as often formalized by (i.e., as abstracted into a mathematical model by) a **cumulative (probability) distribution function** (cdf) (or by other, related functions which we'll meet below).

- E.g., though we may not know our blood cholesterol levels before we observed them, histograms of such measurements tend to have a typical form, perhaps “normal-looking”.
- Or, as we have discussed in Lecture Chapter 1, we may have good reason to assume that the CLT is operating to give approximate normality.

**Definition 3.1** (cdf).

- *The cumulative (probability) distribution function of a random variable  $Y$  is given by*

$$F(y) = P(Y \leq y),$$

*where  $P(Y \leq y)$  is interpreted as the “probability” of the random variable  $Y$  being less than or equal to some number  $y$ .*

- *Note the lowercase  $y$  represents a fixed (non-random) that must be specified in order to get a value for the function  $F$ , just like a typical mathematical function.*
- *This definition seems to imply that the cdf,  $F$ , is defined as a function of some sort of probability function,  $P$ , as if  $P$  exists first. In practice, we typically specify the cdf,  $F$ , (or pmf/pdf, below), which induces a corresponding probability function  $P$ , which we do not discuss. As mentioned in §1.2, we avoid a formal definition of probability.*

**Definition 3.2** (Continuous Random Variable).  *$Y$  is said to be a **continuous random variable** if its cdf is a continuous function.*

- We may use continuous rvs to model measurements such as height, weight, area, volume, etc.
- Intuitively, but somewhat loosely speaking, a continuous random variable can (conceptually) assume any values in an interval.
- Also, intuitively, a continuous random variable has positive probability of being in an interval (of positive length), no matter how small (except in cases such as negative height, etc.).
- Somewhat conversely,  $P(Y = y) = 0$  for a continuous random variable,  $Y$ .

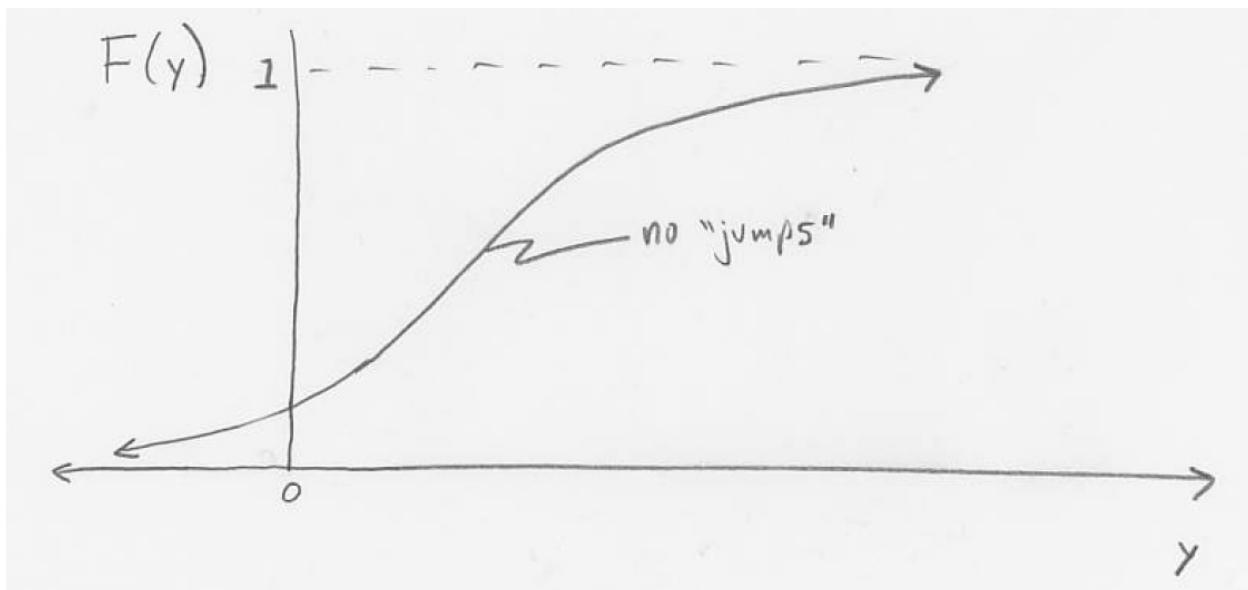


Figure 3.1: Continuous cdf.

**Example 3.1** (Normal (Gaussian) Random Variable).

- We write

$$Y \sim N(\mu, \sigma^2)$$

as short-hand notation for a random variable  $Y$  with a normal distribution with parameters  $-\infty < \mu < \infty$  and  $\sigma > 0$ .

- Let

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y-\mu}{\sigma}\right)^2\right) \quad -\infty < y < \infty.$$

- Then, the cdf of  $Y$  is given by

$$F(y) = \int_{-\infty}^y f(t) dt.$$

- We can loosely think of  $F(y)$  as “summing” probability to the left of some value,  $y$ . Again, note use of lowercase  $y$  as fixed (but unspecified) value.
- See Figure 3.3 ([KNNL05, Table B.1]) for a standard normal (i.e.,  $N(0, 1)$ ) “cdf table,” i.e., a “z-table.” You should already know how to use a z-table!!! We will often choose to use R, however, to compute normal (or other) probabilities for us.

- We will use normal random variables as models for observed data for much of this course, with exception of the binomial and Poisson distributions for logistic or Poisson regression, respectively, if we have enough time.
- We will also use several other distributions, including some that are related to the normal distribution ( $\chi^2$ ,  $t$ ,  $F$ ) or, for Bayesian analyses, (prior/posterior) distributions for the parameters of distributions

(e.g., normal,  $t$ , gamma, inverse gamma (or scaled inverse  $\chi^2$ ), beta, Dirichlet, Wishart).

The cdf is not the only way to describe a random variable.

**Definition 3.3** (Probability Density Function (pdf)).

- If there exists a function  $f(y)$  such that

$$F(y) = \int_{-\infty}^y f(t) dt$$

is a cdf, then  $f(y)$  is called a **probability density function (pdf)**.

- Note that this definition is for a generic continuous random variable.
- In Example 3.1,  $f(y)$  is the pdf of a normal random variable with parameters (mean)  $\mu$  and (standard deviation)  $\sigma$ .
- $\phi$  and  $\Phi$  are often reserved to denote the pdf and cdf, respectively, of the standard normal distribution.

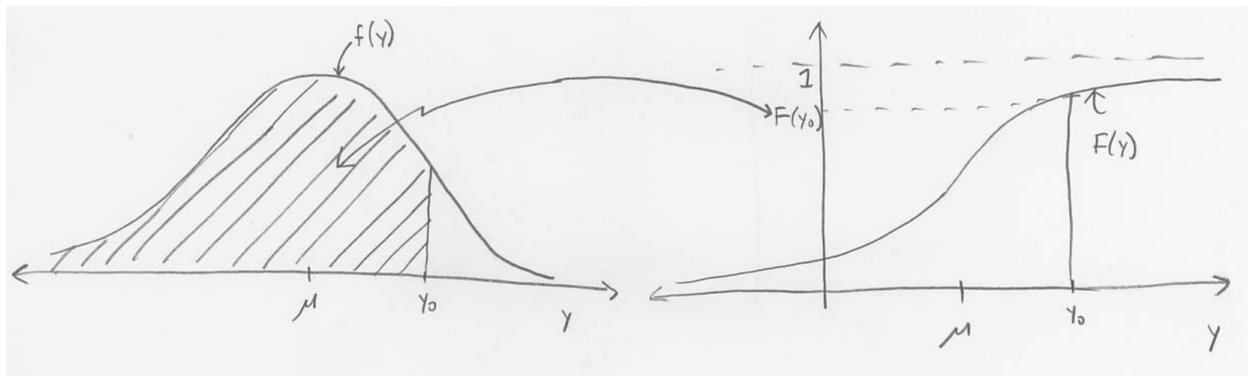


Figure 3.2: Relationship between pdf and cdf.

**Example 3.2** (Z–Table). Use Figure 3.3 ([KNNL05, Table B.1]) to obtain  $P(Z \leq 1.96)$  where  $Z$  is a standard normal rv (my shorthand for “random variable”).

The following chunk explores some of R’s capabilities for computing normal probabilities. Compare to Figure 3.3.

```
> ### Plot  $N(0, 1)$  pdf (often denoted lowercase phi)
> curve(dnorm(x, mean=0, sd=1), from=-3, to = 3,
+         ylab="f(y)", xlab="y")
```

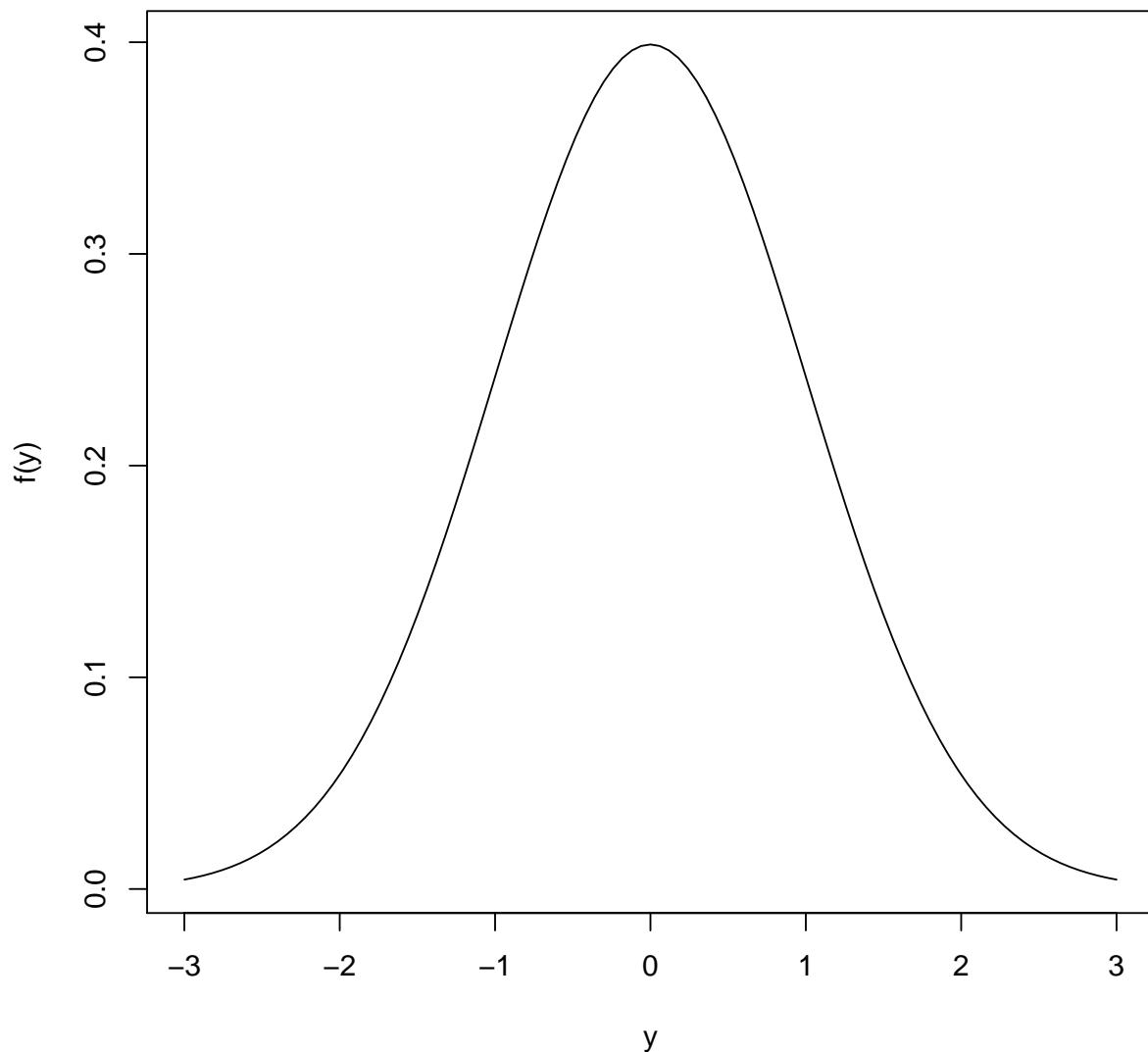
TABLE B.1 Cumulative Probabilities of the Standard Normal Distribution.

| $z$ | Entry is area $A$ under the standard normal curve from $-\infty$ to $z(A)$ |       |       |       |       |       |       |       |       |       |
|-----|--|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|     | .00  | .01   | .02   | .03   | .04   | .05   | .06   | .07   | .08   | .09   |
| .0  | .5000  | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| .1  | .5398  | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| .2  | .5793  | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| .3  | .6179  | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| .4  | .6554  | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |
| .5  | .6915  | .6950 | .6985 | .7019 | .7054 | .7088 | .7123 | .7157 | .7190 | .7224 |
| .6  | .7257  | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| .7  | .7580  | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| .8  | .7881  | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| .9  | .8159  | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| 1.0 | .8413  | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| 1.1 | .8643  | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| 1.2 | .8849  | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | .9015 |
| 1.3 | .9032  | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |
| 1.4 | .9192  | .9207 | .9222 | .9236 | .9251 | .9265 | .9279 | .9292 | .9306 | .9319 |
| 1.5 | .9332  | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 | .9418 | .9429 | .9441 |
| 1.6 | .9452  | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 | .9525 | .9535 | .9545 |
| 1.7 | .9554  | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 | .9616 | .9625 | .9633 |
| 1.8 | .9641  | .9649 | .9656 | .9664 | .9671 | .9678 | .9686 | .9693 | .9699 | .9706 |
| 1.9 | .9713  | .9719 | .9726 | .9732 | .9738 | .9744 | .9750 | .9756 | .9761 | .9767 |
| 2.0 | .9772  | .9778 | .9783 | .9788 | .9793 | .9798 | .9803 | .9808 | .9812 | .9817 |
| 2.1 | .9821  | .9826 | .9830 | .9834 | .9838 | .9842 | .9846 | .9850 | .9854 | .9857 |
| 2.2 | .9861  | .9864 | .9868 | .9871 | .9875 | .9878 | .9881 | .9884 | .9887 | .9890 |
| 2.3 | .9893  | .9896 | .9898 | .9901 | .9904 | .9906 | .9909 | .9911 | .9913 | .9916 |
| 2.4 | .9918  | .9920 | .9922 | .9925 | .9927 | .9929 | .9931 | .9932 | .9934 | .9936 |
| 2.5 | .9938  | .9940 | .9941 | .9943 | .9945 | .9946 | .9948 | .9949 | .9951 | .9952 |
| 2.6 | .9953  | .9955 | .9956 | .9957 | .9959 | .9960 | .9961 | .9962 | .9963 | .9964 |
| 2.7 | .9965  | .9966 | .9967 | .9968 | .9969 | .9970 | .9971 | .9972 | .9973 | .9974 |
| 2.8 | .9974  | .9975 | .9976 | .9977 | .9977 | .9978 | .9979 | .9979 | .9980 | .9981 |
| 2.9 | .9981  | .9982 | .9982 | .9983 | .9984 | .9984 | .9985 | .9985 | .9986 | .9986 |
| 3.0 | .9987  | .9987 | .9987 | .9988 | .9988 | .9989 | .9989 | .9989 | .9990 | .9990 |
| 3.1 | .9990  | .9991 | .9991 | .9991 | .9992 | .9992 | .9992 | .9992 | .9993 | .9993 |
| 3.2 | .9993  | .9993 | .9994 | .9994 | .9994 | .9994 | .9994 | .9995 | .9995 | .9995 |
| 3.3 | .9995  | .9995 | .9995 | .9996 | .9996 | .9996 | .9996 | .9996 | .9996 | .9997 |
| 3.4 | .9997  | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9998 |

| Selected Percentiles         |       |       |       |       |       |       |       |
|------------------------------|-------|-------|-------|-------|-------|-------|-------|
| Cumulative probability $A$ : | .90   | .95   | .975  | .98   | .99   | .995  | .999  |
| $z(A)$ :                     | 1.282 | 1.645 | 1.960 | 2.054 | 2.326 | 2.576 | 3.090 |

Figure 3.3: Z-table (source: [KNNL05, Table B.1]).



```
> ### cdf F(1.96) the hard way (often denoted uppercase Phi)
> integrate(dnorm,lower=-Inf, upper=1.96, mean=0, sd=1)

0.9750021 with absolute error < 1.3e-06

> ### cdf F(1.96) the easy way (often denoted uppercase Phi)
> pnorm(1.96)

[1] 0.9750021
```

**Definition 3.4** (Discrete Random Variable).

- $Y$  is said to be a **discrete random variable** if its cdf is a step function.
- Intuitively, a discrete (categorical, factor) random variable can assume only a countable, perhaps finite, number of values (with positive probability).

**Example 3.3** (Binomial Random Variable).

- Let  $Y$  be an rv that counts the number of heads in  $n$  tosses of a coin.
- $Y$  is often specified as a binomial random variable (along with additional assumptions), denoted  $\text{binom}(n, p)$ , with cdf,

$$\begin{aligned} F(y) &= P(Y \leq y) = \sum_{i \leq y} P(Y = i) \\ &= \sum_{i \leq y} \binom{n}{i} p^i (1-p)^{(n-i)}, \end{aligned}$$

where summation occurs over integers in  $[0, y]$ .  $n$  and  $0 < p < 1$  are the parameters of the binomial cdf ( $n$  is an integer greater than or equal to one).

- In this discrete rv case,  $F(y)$  really does sum probability to the left of (and including)  $y$ . (Don't let recycled notation confuse you.)

The pmf is another way to describe a discrete random variable.

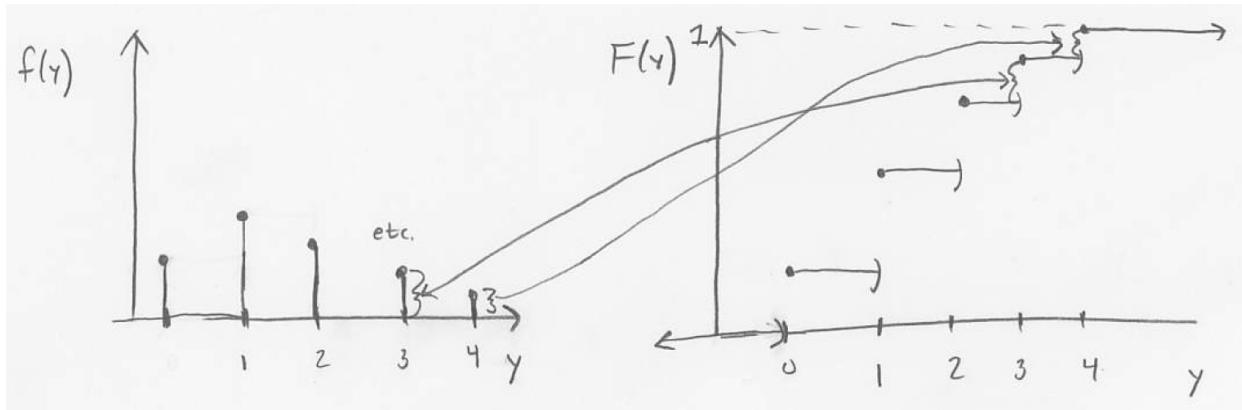


Figure 3.4: Relationship between pmf and cdf.

**Definition 3.5** (Probability Mass Function (pmf)). *If*

$$f(y) = P(Y = y),$$

*then  $f(y)$  is called the **probability mass function (pmf)**.*

We may use discrete rvs to model measurements such as counts (Poisson regression), presence/absence (logistic regression), names, categories, etc.

**Example 3.4** (Binomial Random Variable (cont'd)).

- The **probability mass function (pmf)** of a binomial random variable  $\text{binom}(n, p)$  is given by

$$f(y) = P(Y = y) = \begin{cases} \binom{n}{y} p^y (1-p)^{(n-y)} & y = 0, 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

- A binomial rv is often used to model  $n$  (assumed) independent “0/1” or “success/failure” (Bernoulli) types of measurements with  $a(n)$  (as-

*sumed) constant probability of “success”,  $p$ ; e.g., presence/absence of bird nests in  $n$  trees.*

- *It is fundamental to **logistic regression** ([Wak13, Chap. 7]), which we may cover, time permitting.*

### Example 3.5 (Tossing a Coin).

- *What’s the probability of getting two heads in five (independent) tosses of a fair coin?*
- *Use the “binomial table” given nearby. (Or use R.)*
- *(Incidentally, look at  $P(Y \leq 4)$  for  $Y \sim \text{binom}(5, 0.5)$  in the table; this could be intentional...more in class.)*

The following Chunk explores some of R’s capabilities for computing binomial probabilities.

```
> ### cdf F(2) (for tossing fair coin 5 times (binom(n=5,p=0.5))
> pbinom(2,size=5,prob=0.5)

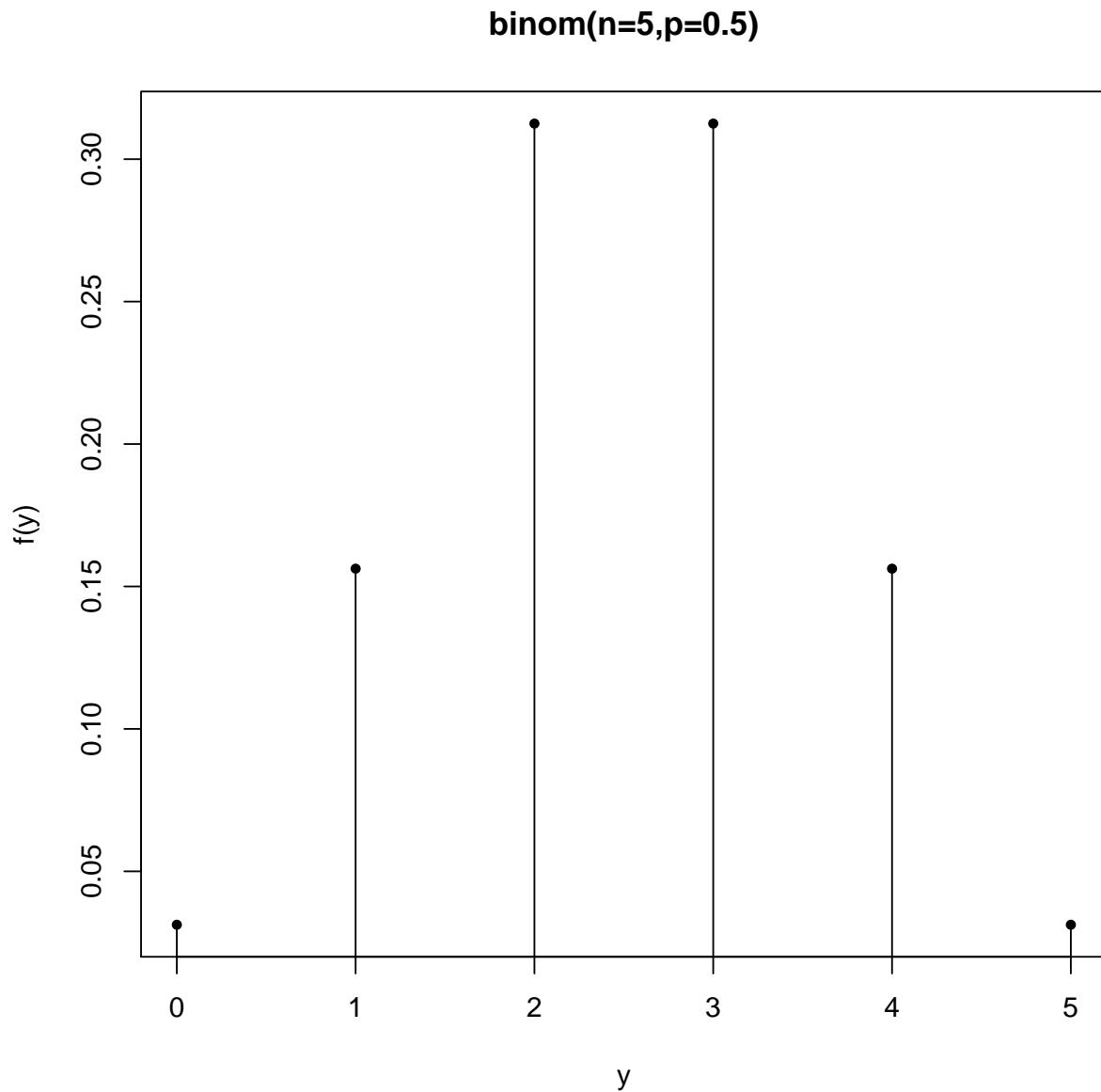
[1] 0.5

> ### binomial(n=5,p=0.5)
> plot(0:5, binprobs<- dbinom(0:5,size=5,p=0.5),
+       ylab="f(y)", xlab="y",
+       pch=20, main="binom(n=5,p=0.5)")
> segments(0:5,rep(0,5),0:5,binprobs)
```

Table II Cumulative Binomial Probabilities  $P(X \leq x)$ 

| $n$ | $x$ | $P$    |        |        |        |        |        |        |        |        |        |               |
|-----|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------------|
|     |     | 0.1    | 0.2    | 0.3    | 0.4    | 0.5    | 0.6    | 0.7    | 0.8    | 0.9    | 0.95   | 0.99          |
| 1   | 0   | 0.9000 | 0.8000 | 0.7000 | 0.6000 | 0.5000 | 0.4000 | 0.3000 | 0.2000 | 0.1000 | 0.0500 | 0.0100        |
|     | 2   | 0      | 0.8100 | 0.6400 | 0.4900 | 0.3600 | 0.2500 | 0.1600 | 0.0900 | 0.0400 | 0.0100 | 0.0025 0.0001 |
| 3   | 1   | 0.9900 | 0.9600 | 0.9100 | 0.8400 | 0.7500 | 0.6400 | 0.5100 | 0.3600 | 0.1900 | 0.0975 | 0.0199        |
|     | 0   | 0.7290 | 0.5120 | 0.3430 | 0.2160 | 0.1250 | 0.0640 | 0.0270 | 0.0080 | 0.0010 | 0.0001 | 0.0000        |
| 4   | 1   | 0.9720 | 0.8960 | 0.7840 | 0.6480 | 0.5000 | 0.3520 | 0.2160 | 0.1040 | 0.0280 | 0.0073 | 0.0003        |
|     | 2   | 0.9990 | 0.9920 | 0.9730 | 0.9360 | 0.8750 | 0.7840 | 0.6570 | 0.4880 | 0.2710 | 0.1426 | 0.0297        |
| 5   | 0   | 0.6561 | 0.4096 | 0.2401 | 0.1296 | 0.0625 | 0.0256 | 0.0081 | 0.0016 | 0.0001 | 0.0000 | 0.0000        |
|     | 1   | 0.9477 | 0.8192 | 0.6517 | 0.4752 | 0.3125 | 0.1792 | 0.0837 | 0.0272 | 0.0037 | 0.0005 | 0.0000        |
| 6   | 2   | 0.9963 | 0.9728 | 0.9163 | 0.8208 | 0.6875 | 0.5248 | 0.3483 | 0.1808 | 0.0523 | 0.0140 | 0.0006        |
|     | 3   | 0.9999 | 0.9984 | 0.9919 | 0.9744 | 0.9375 | 0.8704 | 0.7599 | 0.5904 | 0.3439 | 0.1855 | 0.0394        |
| 7   | 0   | 0.5905 | 0.3277 | 0.1681 | 0.0778 | 0.0313 | 0.0102 | 0.0024 | 0.0003 | 0.0000 | 0.0000 | 0.0000        |
|     | 1   | 0.9185 | 0.7373 | 0.5282 | 0.3370 | 0.1875 | 0.0870 | 0.0308 | 0.0067 | 0.0005 | 0.0000 | 0.0000        |
| 8   | 2   | 0.9914 | 0.9421 | 0.8369 | 0.6826 | 0.5000 | 0.3174 | 0.1631 | 0.0579 | 0.0086 | 0.0012 | 0.0000        |
|     | 3   | 0.9995 | 0.9933 | 0.9692 | 0.9130 | 0.8125 | 0.6630 | 0.4718 | 0.2627 | 0.0815 | 0.0226 | 0.0010        |
| 9   | 4   | 1.0000 | 0.9997 | 0.9976 | 0.9898 | 0.6988 | 0.9222 | 0.8319 | 0.6723 | 0.4095 | 0.2262 | 0.0490        |
|     | 0   | 0.5314 | 0.2621 | 0.1176 | 0.0467 | 0.0156 | 0.0041 | 0.0007 | 0.0001 | 0.0000 | 0.0000 | 0.0000        |
| 10  | 1   | 0.8857 | 0.6554 | 0.4202 | 0.2333 | 0.1094 | 0.0410 | 0.0109 | 0.0016 | 0.0001 | 0.0000 | 0.0000        |
|     | 2   | 0.9842 | 0.9011 | 0.7443 | 0.5443 | 0.3438 | 0.1792 | 0.0705 | 0.0170 | 0.0013 | 0.0001 | 0.0000        |
| 11  | 3   | 0.9987 | 0.9830 | 0.9295 | 0.8208 | 0.6563 | 0.4557 | 0.2557 | 0.0989 | 0.0159 | 0.0022 | 0.0000        |
|     | 4   | 0.9999 | 0.9984 | 0.9891 | 0.9590 | 0.9806 | 0.7667 | 0.5798 | 0.3446 | 0.1143 | 0.0328 | 0.0015        |
| 12  | 5   | 1.0000 | 0.9999 | 0.9993 | 0.9959 | 0.9844 | 0.9533 | 0.8824 | 0.7379 | 0.4686 | 0.2649 | 0.0585        |
|     | 0   | 0.4783 | 0.2097 | 0.0824 | 0.0280 | 0.0078 | 0.0016 | 0.0002 | 0.0000 | 0.0000 | 0.0000 | 0.0000        |
| 13  | 1   | 0.8503 | 0.5767 | 0.3294 | 0.1586 | 0.0625 | 0.0188 | 0.0038 | 0.0004 | 0.0000 | 0.0000 | 0.0000        |
|     | 2   | 0.9743 | 0.8520 | 0.6471 | 0.4199 | 0.2266 | 0.0963 | 0.0288 | 0.0047 | 0.0002 | 0.0000 | 0.0000        |
| 14  | 3   | 0.9973 | 0.9667 | 0.8740 | 0.7102 | 0.5000 | 0.2898 | 0.1260 | 0.0333 | 0.0027 | 0.0002 | 0.0000        |
|     | 4   | 0.9998 | 0.9953 | 0.9712 | 0.9037 | 0.7734 | 0.5801 | 0.3529 | 0.1480 | 0.0257 | 0.0038 | 0.0000        |
| 15  | 5   | 1.0000 | 0.9996 | 0.9962 | 0.9812 | 0.9375 | 0.8414 | 0.6706 | 0.4233 | 0.1497 | 0.0444 | 0.0020        |
|     | 6   | 1.0000 | 1.0000 | 0.9998 | 0.9984 | 0.9922 | 0.9720 | 0.9176 | 0.7903 | 0.5217 | 0.3017 | 0.0679        |
| 16  | 0   | 0.4305 | 0.1678 | 0.0576 | 0.0168 | 0.0039 | 0.0007 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000        |
|     | 1   | 0.8131 | 0.5033 | 0.2553 | 0.1064 | 0.0352 | 0.0085 | 0.0013 | 0.0001 | 0.0000 | 0.0000 | 0.0000        |
| 17  | 2   | 0.9619 | 0.7969 | 0.5518 | 0.3154 | 0.1445 | 0.0498 | 0.0113 | 0.0012 | 0.0000 | 0.0000 | 0.0000        |
|     | 3   | 0.9950 | 0.9437 | 0.8059 | 0.5941 | 0.3633 | 0.1737 | 0.0580 | 0.0104 | 0.0004 | 0.0000 | 0.0000        |
| 18  | 4   | 0.9996 | 0.9896 | 0.9420 | 0.8263 | 0.6367 | 0.4059 | 0.1941 | 0.0563 | 0.0050 | 0.0004 | 0.0000        |
|     | 5   | 1.0000 | 0.9988 | 0.9887 | 0.9502 | 0.8555 | 0.6846 | 0.4482 | 0.2031 | 0.0381 | 0.0058 | 0.0001        |
| 19  | 6   | 1.0000 | 0.9999 | 0.9987 | 0.9915 | 0.9648 | 0.8936 | 0.7447 | 0.4967 | 0.1869 | 0.0572 | 0.0027        |
|     | 7   | 1.0000 | 1.0000 | 0.9999 | 0.9993 | 0.9961 | 0.9832 | 0.9424 | 0.8322 | 0.5695 | 0.3366 | 0.0773        |
| 20  | 0   | 0.3874 | 0.1342 | 0.0404 | 0.0101 | 0.0020 | 0.0003 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000        |
|     | 1   | 0.7748 | 0.4362 | 0.1960 | 0.0705 | 0.0195 | 0.0038 | 0.0004 | 0.0000 | 0.0000 | 0.0000 | 0.0000        |
| 21  | 2   | 0.9470 | 0.7382 | 0.4628 | 0.2318 | 0.0889 | 0.0250 | 0.0043 | 0.0003 | 0.0000 | 0.0000 | 0.0000        |
|     | 3   | 0.9917 | 0.9144 | 0.7297 | 0.4826 | 0.2539 | 0.0994 | 0.0253 | 0.0031 | 0.0001 | 0.0000 | 0.0000        |
| 22  | 4   | 0.9991 | 0.9804 | 0.9012 | 0.7334 | 0.5000 | 0.2666 | 0.0988 | 0.0196 | 0.0009 | 0.0000 | 0.0000        |
|     | 5   | 0.9999 | 0.9969 | 0.9747 | 0.9006 | 0.7461 | 0.5174 | 0.2703 | 0.0856 | 0.0083 | 0.0006 | 0.0000        |
| 23  | 6   | 1.0000 | 0.9997 | 0.9957 | 0.9750 | 0.9102 | 0.7682 | 0.5372 | 0.2618 | 0.0530 | 0.0084 | 0.0001        |
|     | 7   | 1.0000 | 1.0000 | 0.9996 | 0.9962 | 0.9805 | 0.9295 | 0.8040 | 0.5638 | 0.2252 | 0.0712 | 0.0034        |
| 24  | 8   | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9980 | 0.9899 | 0.9596 | 0.8658 | 0.6126 | 0.3698 | 0.0865        |

Figure 3.5: Table of cumulative binomial probabilities (Source: forgotten!).



```
> ### Incidentally (see the nearby table)
> pbinom(4,5,0.5)

[1] 0.96875
```

**Example 3.6** (Poisson RV).

- $Y \sim \text{Pois}(\lambda)$ ,  $\lambda > 0$ , with pmf

$$f(y) = \begin{cases} \frac{\lambda^y \exp(-\lambda)}{y!} & y = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

- A Poisson random variable is often used to model count data and is fundamental to **Poisson regression** (aka, log-linear regression; [Wak13, Chap. 6]), which we may get to, if we have time.

**Definition 3.6** (Support of an RV).

- The support of an rv is the set of values for  $y$  where  $f(y) > 0$ , where  $f$  is a pdf, for a continuous rv, or pmf, for a discrete rv.
- We'll sometimes use  $S$  to denote support or  $\mathcal{X}$  for the support of a random variable  $X$  or  $\mathcal{Y}$  for the support of a random variable  $Y$ , etc.

What's the support of a normal random variable?  $\text{binom}(n = 5, p = 0.2)$ ?  $\text{binom}(1, 0.5)$ ?  $\text{Pois}(\lambda)$ ?

**NOTE:** Once we get to traditional “linear regression” and “ANOVA” (both special cases of our linear model), we will be working mostly with Gaussian random variables to model observed data, with exception of logistic regression and Poisson regression and some methods that relax such fully specified probability models by only requiring the specification of a mean (regression function) model and variance model.

**MODELING REMINDER:** We use rvs and their distributions, usually through parameters, as models (mathematical abstractions) of reality. More particularly, we will assume that data has arisen from a random variable with some distribution, and our job is to use the data to estimate (the unknown parameters in) the distribution. So far in this lecture set we haven't talked about data!!! The means, variances, and covariances below are model (or “(super)population” or “theoretical”) quantities, not (“sample”) data quantities (though we could assign a pmf with  $1/n$  probability for each datum of a data set of size  $n$  then use the material below to summarize data, too).

### 3.3 Characteristics of Random Variables

A random variable is fully specified by its distribution (cdf, pdf, pmf). But, we often do not deal directly with an rv's distribution but with some summarizing—and easier to understand—property of the distribution, like the (“population”) **mean, variance, or covariance**. As mentioned, above, we hope to get to methods that rely only on the specification of mean (regression function) and (co-)variance models.

#### 3.3.1 Expected (Mean) Value

**Definition 3.7** (Expected (Mean) Value of a Discrete RV).

- 

$$\mu(Y) = E(Y) = \sum_{y \in S} yP(Y = y) = \sum_{y \in S} yf(y),$$

where  $S$  is the support of the rv  $Y$ .

- *The mean is a measure of center or central tendency of an rv in the sense of being the balance point of the probability masses of the pmf.*

- *Appealing for unimodal and symmetric distributions.*
- *A weighted average of  $y$  values with weights given by the pmf,  $f(y)$ .*

**NOTE:** Use of the expectation (mean) operation notation,  $E$ , is common. We may also use the notation,  $\mu$ , perhaps as a function of covariates and parameters, to denote the expected value of a random variable.

**Example 3.7** (Mean of Discrete Uniform RV).

$$f(y) = \begin{cases} \frac{1}{3} & y \in \overbrace{\{1, 2, 3\}}^S \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} E(Y) &= \sum_{y \in \{1, 2, 3\}} y f(y) = 1(1/3) + 2(1/3) + 3(1/3) \\ &= 1/3 \sum_{y=1}^3 y = \frac{6}{3} = 2 \end{aligned}$$

**Example 3.8** (Mean of a Bernoulli (i.e.,  $\text{bin}(n = 1, p)$ )).

$$\begin{aligned} E(Y) &= \sum_{y \in \{0, 1\}} y f(y) = 0f(0) + 1f(1) \\ &= 0p^0(1-p)^{1-0} + 1p^1(1-p)^{1-1} = 0 + 1p = p \end{aligned}$$

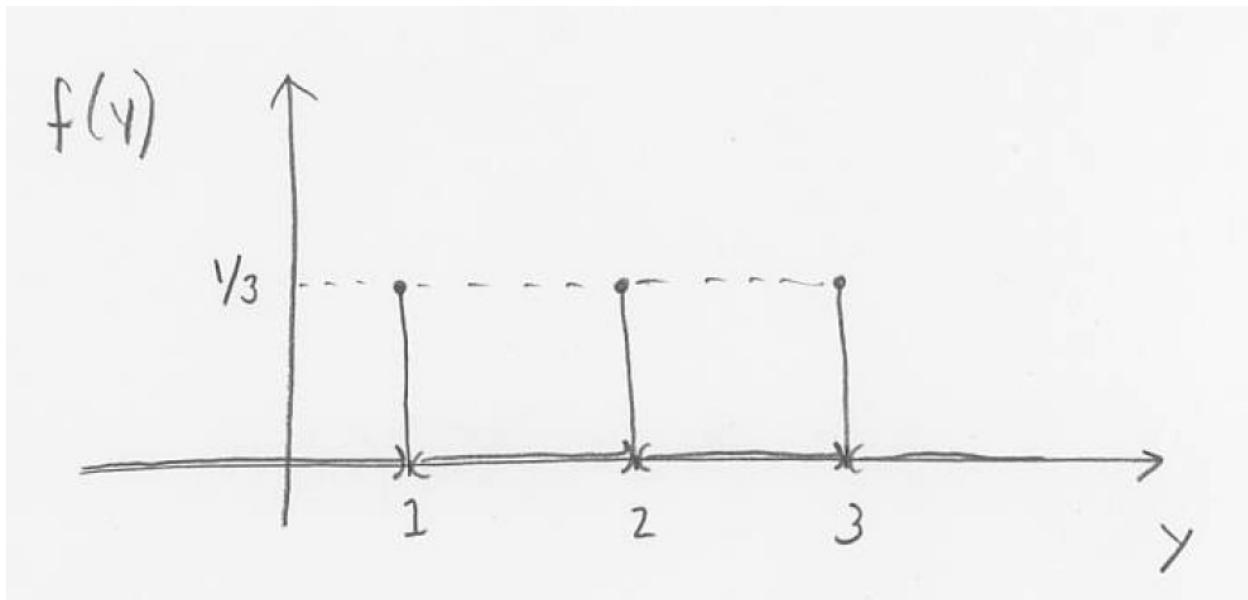


Figure 3.6: Discrete uniform pmf supported on 1, 2, 3.

**Definition 3.8** (Expected (Mean) Value of a Continuous RV).

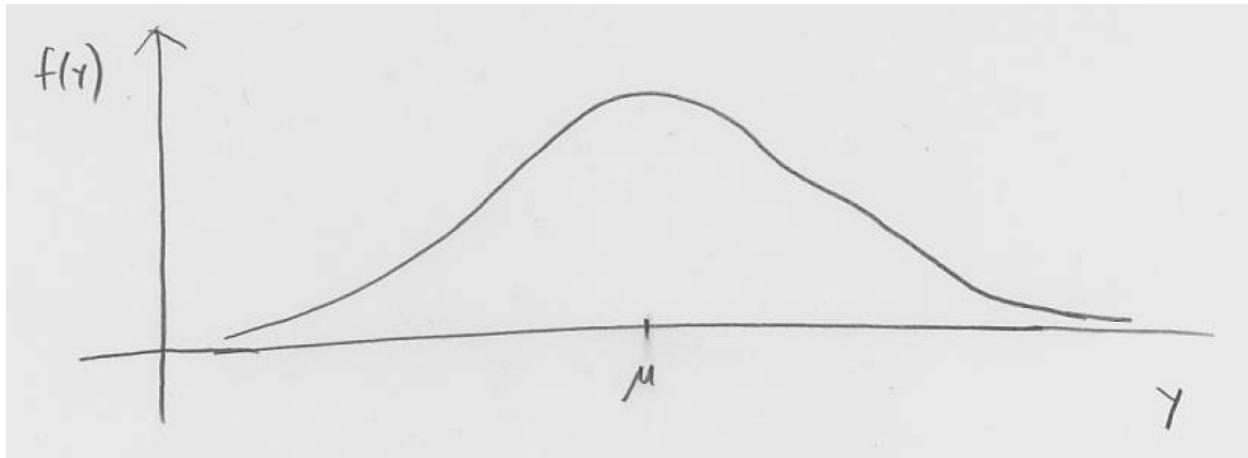
$$E(Y) = \int_{-\infty}^{\infty} yf(y) dy$$

**Example 3.9** (Mean of  $N(\mu, \sigma^2)$ ). If  $Y \sim N(\mu, \sigma^2)$ , then

$$E(Y) = \mu$$

(details omitted).

**Remark 3.1** (Linearity Property of the Expectation Operator).

Figure 3.7: pdf is balanced on mean ( $\mu$ ).

- Let  $a$ ,  $b$ , and  $c$  be arbitrary constants.
- Let  $X$  and  $Y$  be arbitrary rvs whose expectations (expected values, means) exist.
- Then

$$\begin{aligned} E(a + bX + cY) &= E(a) + E(bX) + E(cY) \\ &= a + bE(X) + cE(Y) \end{aligned}$$

- This property follows from properties of (convergent) sums/series (for discrete rvs) or integrals (for continuous rvs). Details omitted.
- This is a particular example of a **linear combination** of rvs with coefficients  $a$ ,  $b$ , and  $c$ . ( $1$  (multiplying  $a$ ) may be thought of as a degenerate rv). We will talk a bit more about linear combinations later in the course when comparing parameters in regression or ANOVA models.
- In words, “the expectation (mean) of the linear combination is the linear combination of the expectations (means).”

**Example 3.10** (Mean of a Difference).

- Let  $X \sim N(\mu_X, \sigma_X^2)$   $Y \sim N(\mu_Y, \sigma_Y^2)$ .
- Then,

$$E(X - Y) = \mu_X - \mu_Y.$$

Note  $a = 0$ ,  $b = 1$  and  $c = -1$  in the linearity property just presented, and we have introduced subscripts on our  $\mu$  notation for expectations.

- “The mean of the difference is the difference of the means.”

### 3.3.2 Variance Operator

**Definition 3.9** (Variance of an RV).

- 

$$\begin{aligned} \text{Var}(Y) &= E(Y - E(Y))^2 \\ &= E(Y^2) - (E(Y))^2 \end{aligned}$$

- The variance of an rv is a **measure of the spread or dispersion** of the possible values of an rv.
- May be thought of as the (weighted) **average squared deviation** of a rv from its mean.
- (Let’s not use “mean squared error” until we learn about bias.)

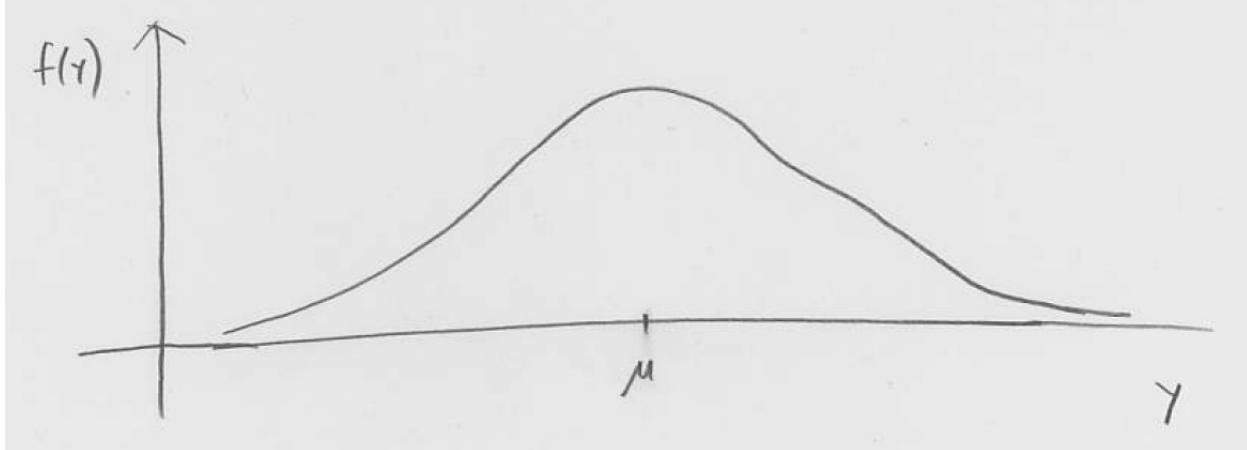
**Example 3.11** (Variance of  $N(\mu, \sigma^2)$ ).

- If  $Y \sim N(\mu, \sigma^2)$ ,

- then

$$\text{Var}(Y) = \int_{-\infty}^{\infty} (y - E(Y))^2 f(y) dy = \sigma^2$$

(details omitted).



**Example 3.12** (Variance of A Bernoulli (i.e.,  $\text{binom}(n = 1, p)$ )).

$$\begin{aligned}\text{Var}(Y) &= \sum_{y=0}^1 (y - p)^2 p^y (1 - p)^{1-y} \\ &= (0 - p)^2 p^0 (1 - p)^{1-0} + (1 - p)^2 p^1 (1 - p)^{1-1} \\ &= p^2 (1 - p) + (1 - p)^2 p \\ &= p(1 - p)(p + (1 - p)) \\ &= p(1 - p)\end{aligned}$$

**Definition 3.10** (Standard Deviation of an RV).

$$SD(Y) = \sqrt{\text{Var}(Y)}$$

### Useful Property of Variance Operator

$$\text{Var}(a + bY) = b^2 \text{Var}(Y)$$

**Example 3.13.** Let  $Y \sim N(\mu, \sigma^2)$  and  $Z = \frac{Y-\mu}{\sigma}$

$$\begin{aligned} \text{Var}(Z) &= \text{Var}\left(-\frac{\mu}{\sigma} + \frac{1}{\sigma}Y\right) \\ &= \frac{1}{\sigma^2} \text{Var}(Y) \\ &= \frac{\sigma^2}{\sigma^2} = 1 \end{aligned}$$

## 3.4 Random Vectors

Here, we treat only a few fundamental concepts involving groups of random variables. We will treat matrices and vectors formally in a subsequent chapter of our notes before giving more details on random vectors.

### 3.4.1 Covariance Operator & Its Properties

Covariance is a property of two random variables.

**Definition 3.11** (Covariance).

$$\begin{aligned} \text{Cov}(X, Y) &= E(X - E(X))(Y - E(Y)) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

**Example 3.14** (Covariance of Bivariate Normal).

$$f(x, y) = \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2(1-\rho)^2}} \times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \left( \frac{x-\mu_X}{\sigma_X} \right)^2 + \left( \frac{y-\mu_Y}{\sigma_Y} \right)^2 - 2\rho \left( \frac{x-\mu_X}{\sigma_X} \right) \left( \frac{y-\mu_Y}{\sigma_Y} \right) \right] \right\}$$

$$\begin{aligned} Cov(X, Y) &= \int \int (x - \mu_X)(y - \mu_Y) f(x, y) dx dy \\ &= \sigma_X \sigma_Y \rho \end{aligned}$$

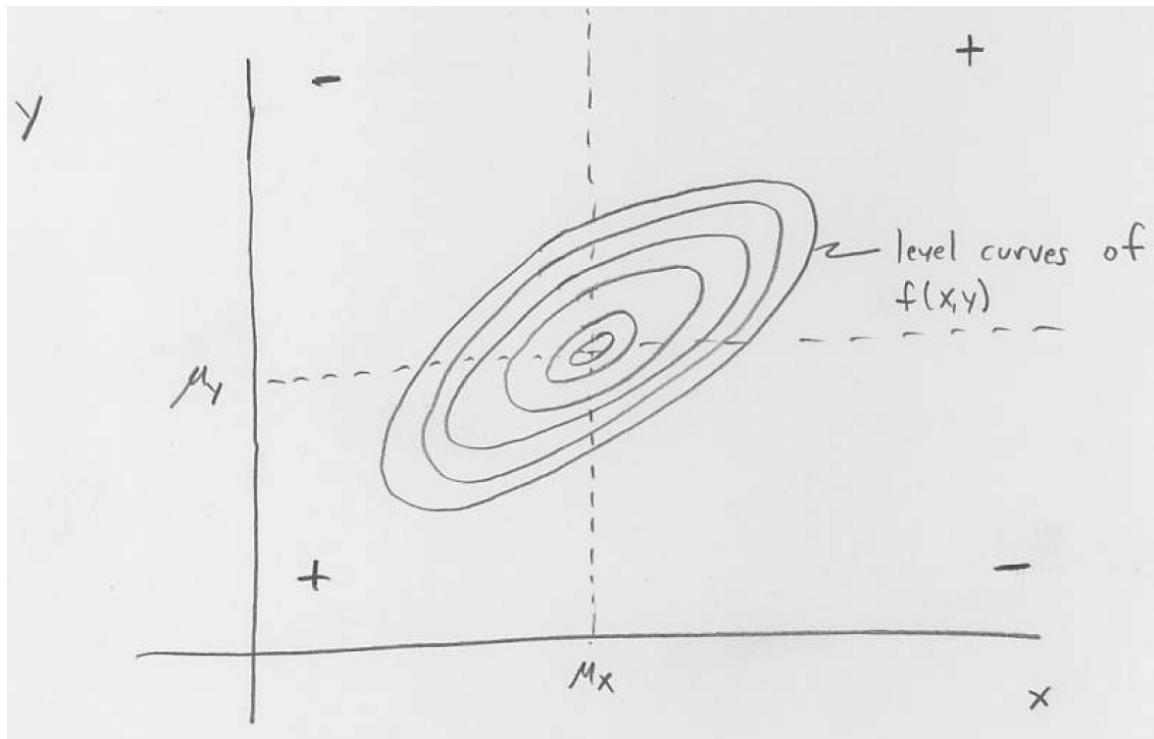


Figure 3.8: Sketch to illustrate covariance.

**Remark 3.2** (Properties of Variance & Covariance Operators).

$$\text{Cov}(Y, Y) = \text{Var}(Y) \quad (\text{right}?!)$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^n c_i Y_i\right) &= \sum_{i=1}^n \sum_{j=1}^n c_i c_j \text{Cov}(Y_i, Y_j) \\ &= \sum_{i=1}^n c_i^2 \text{Var}(Y_i) + 2 \sum_{i < j}^n c_i c_j \text{Cov}(Y_i, Y_j) \end{aligned}$$

Take  $n = 2$  &  $c_1 = c_2 = 1$  to get result for  $\text{Var}(X + Y)$  above.

$$\text{Var}(a) = 0$$

$$\text{Cov}(a, Y) = 0$$

**Definition 3.12** (Correlation). *The correlation between two variables is the covariance between the variables' standardized versions. It's unitless.*

$$\begin{aligned} \text{Cor}(X, Y) &= \text{Cov}\left(\frac{X - E(X)}{\sqrt{\text{Var}(X)}}, \frac{Y - E(Y)}{\sqrt{\text{Var}(Y)}}\right) \\ &= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} \end{aligned}$$

**Example 3.15** (Covariance of Bivariate Normal (cont'd)).

$$f(x, y) = \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2(1-\rho)^2}} \times \\ \exp\left\{-\frac{1}{2(1-\rho^2)} \left[ \left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) \right] \right\}$$

$$\text{Cor}(X, Y) = \rho$$

### 3.4.2 Independence

We will have more to say about independence and joint distributions after covering vectors more formally. For now, we give a brief look at independence.

$X, Y$  are said to be independent if their joint distribution (e.g. cdf, pdf, pmf) factors into the product of marginal distributions, i.e., (using pdf/pmf notation) if

$$f(x, y) = f(x)f(y).$$

(Not same  $f$  of course. Recycling notation!). Less precisely but more intuitively,  $X$  and  $Y$  are independent if the values of  $X$  tell us nothing about the values of  $Y$ , and vice-versa. Always, if  $X$  and  $Y$  are independent, then  $\text{Cov}(X, Y) = 0$ . But, if  $\text{Cov}(X, Y) = 0$ , then  $X$  and  $Y$  are not necessarily independent, normality being an exception: if  $X$  and  $Y$  are normally distributed, then  $\text{Cov}(X, Y) = 0$  implies  $f(x, y) = f(x)f(y)$  (look at the bivariate normal pdf nearby).

**Example 3.16** (Independent Coin Tosses). *Let  $X$  and  $Y$  be the result of two coin tosses. In most circumstances, it would be reasonable to think that the result of one toss has nothing to do with the result of another.*

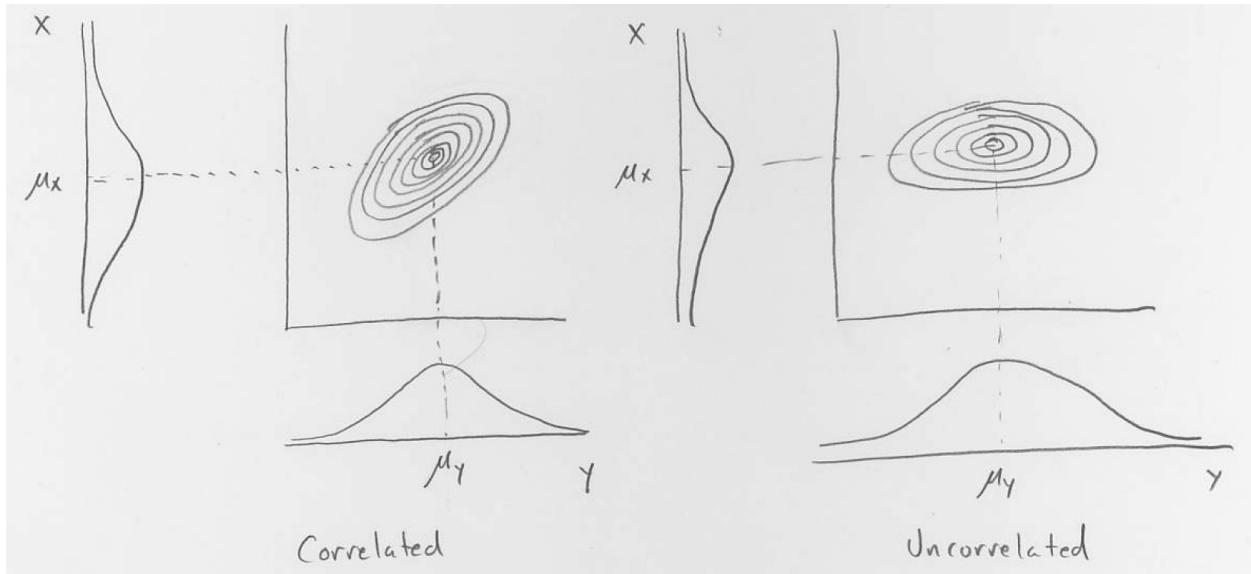


Figure 3.9: Sketch of marginal and joint distributions.

Thus, independence seems eminently reasonable. Assuming a Bernoulli ( $\text{binom}(n = 1, p)$ ) for each toss we have

$$f(x, y) = f(x)f(y) = p_X^x(1 - p_X)^{1-x}p_Y^y(1 - p_Y)^{1-y}.$$

It may be reasonable to assume  $p_X = p_Y = 0.5$ .

**NOTE:** Independence is typically an assumption (hopefully reasonable) made to facilitate model building (rather than something to be checked after a model is already specified). It is often easier to specify  $f(x)$  and  $f(y)$  than to specify  $f(x, y)$  directly.

### 3.5 Central Limit Theorem

We have already touched on the CLT in a previous chapter (1). There are various versions of the CLT. We give one here.

**Theorem 3.1** (A Central Limit Theorem (CLT)). *If  $Y_1, Y_2, \dots, Y_n$  are independent rvs from the same distribution with same mean  $E(Y_i) = \mu$  and  $\text{Var}(Y_i) = \sigma^2 < \infty$ , then*

$$\bar{Y}_n \stackrel{\text{"dot means approx"} \atop \sim}{\overbrace{\cdot}} N(\mu, \sigma^2/n)$$

and the approximation improves with increasing  $n$ .

**NOTE:** If  $Y_i \stackrel{\text{ind}}{\sim} N(\mu, \sigma^2)$ , then  $\bar{Y}_n \sim N(\mu, \sigma^2/n)$ , exactly, for any positive integer  $n$ .

## 3.6 Linear Functions of an RV

This is a particular case of linear combination, in the next section, and we will give the matrix analog in a subsequent chapter of notes.

If  $Y$  is an rv, then  $X = a + bY$  is a **linear function** of  $Y$  (and is also an rv, of course) and has mean

$$E(X) = a + bE(Y)$$

and variance

$$\text{Var}(X) = b^2 \text{Var}(Y).$$

Moreover (this is new), if  $Y$  is normally distributed, then so is  $X$ .

**Example 3.17** (Linear Function of Normal RV). *If  $Y \sim N(\mu, \sigma^2)$ , then  $Z = \frac{Y-\mu}{\sigma} \sim N(0, 1)$ . ( $a = \frac{-\mu}{\sigma}$ ,  $b = \frac{1}{\sigma}$ )*

Of course, we knew this already, right?!

**Example 3.18** (Linear Function of Normal RV (again)). If  $Z \sim N(0, 1)$ , then  $Y = \mu + \sigma Z \sim N(\mu, \sigma^2)$ . ( $a = \mu$ ,  $b = \sigma$ )

### 3.7 Linear Combinations of RVs

We will return to linear combinations of random variables, using matrix notation, in a subsequent chapter. For now, we introduce linear combinations using non-matrix notation.

- If  $Y_i$  are rvs and  $c_i$  are constants,  $i = 1, \dots, n$ , then  $\sum_{i=1}^n c_i Y_i$  is a **linear combination** of the  $Y_i$  with **coefficients**  $c_i$ .
- $\sum_{i=1}^n c_i Y_i$  has mean

$$E\left(\sum_{i=1}^n c_i Y_i\right) = \sum_{i=1}^n c_i E(Y_i).$$

In words, “*the mean of linear combination is the linear combination of means.*”

- Furthermore, if the  $Y_i$  are uncorrelated (e.g., independent), then

$$\text{Var}\left(\sum_{i=1}^n c_i Y_i\right) = \sum_{i=1}^n c_i^2 \text{Var}(Y_i).$$

In words, “*the variance of a linear combination is (almost) the linear combination of variances (with coefficients now squared), if the variables are uncorrelated.*”

- If, in addition, the  $Y_i$  are normally distributed, then so is the linear combination:

$$\sum_{i=1}^n c_i Y_i \sim N\left(\sum_{i=1}^n c_i E(Y_i), \sum_{i=1}^n c_i^2 \text{Var}(Y_i)\right)$$

- We will use similar results on linear combinations a lot. Much more later.



# Lecture 4

## Matrices & Vectors

### Contents

---

|       |  |     |
|-------|--|-----|
| 4.1   | Notation, Dimension, Rows, Columns, Elements . . . . .                 | 117 |
| 4.2   | Matrix Arithmetic . . . . .  | 121 |
| 4.2.1 | Addition/Subtraction . . . . .   | 122 |
| 4.2.2 | Multiply a Matrix by a Scalar . . . . .                                | 124 |
| 4.2.3 | Matrix Multiplication . . . . .  | 124 |
| 4.2.4 | Matrix Transpose . . . . .   | 126 |
| 4.2.5 | Special Matrices . . . . .   | 129 |
| 4.2.6 | Linear Dependence & Rank . . . . .                                     | 132 |
| 4.3   | Combining Things: Random Vectors and Matrices . . . . .                | 141 |
| 4.3.1 | Expectation of a Random Vector/Matrix . . . . .                        | 142 |
| 4.3.2 | Variance(-Covariance) Matrix . . . . .                                 | 142 |
| 4.3.3 | Linearity of Expectation Operator (just as in scalar case) . . . . .   | 143 |
| 4.3.4 | Variance(-Covariance) of $\mathbf{a} + \mathbf{B}\mathbf{Y}$ . . . . . | 143 |
| 4.3.5 | Distribution of Linear Function of Normal RV . . . . .                 | 143 |

---

***Main Objectives:***

- Learn basic matrix operations, including matrix addition, matrix multiplication, scalar multiplication of a matrix, transpose, inverse.
- Continue to familiarize ourselves with R and some of its basic matrix and vector functionality.

- Preview common matrices and vectors used in regression and ANOVA so that these are familiar when we discuss and perform regression and ANOVA computations, per se.
  - Apply previous results (Lecture 3) on random variables to random matrices and random vectors (mostly vectors).
  - Use matrices and vectors to specify a general linear model, of which regression and ANOVA are special cases.
- 
- $\mathcal{O}$

***Additional Reading:***

Much of the material here is based on [KNNL05, Sec. 5.1 – 5.11], which is on reserve in the Cline Library; you are welcome to read it if you feel the need. But, I suggest that these notes, here, are sufficient.

We begin to transition more toward our required reading and include more references to [Wak13].  $\mathcal{R}$

Some familiarity with matrices and vectors is required for a good understanding of the models that underlie regression and ANOVA, or linear statistical models in general.

**Definition 4.1** (Matrix). *A two-dimensional array of numbers.*

**Definition 4.2** (Matrix Size and Dimensions).

- *The size of a matrix is given by its two dimensions, often called “rows” and “columns.”*
- *Similar to the size of a rectangular room given by its vertical and horizontal dimensions.*
- *An  $r \times c$  (size) matrix has  $r$  rows and  $c$  columns.*

## 4.1 Notation, Dimension, Rows, Columns, Elements

**Example 4.1** (A 2 by 2 Matrix).

$$\mathbf{A}_{2 \times 2} = \begin{bmatrix} 1 & 7 \\ 4 & 3 \end{bmatrix}$$

**Example 4.2** (A 2 by 3 Matrix).

$$\mathbf{B}_{2 \times 3} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}$$

OR

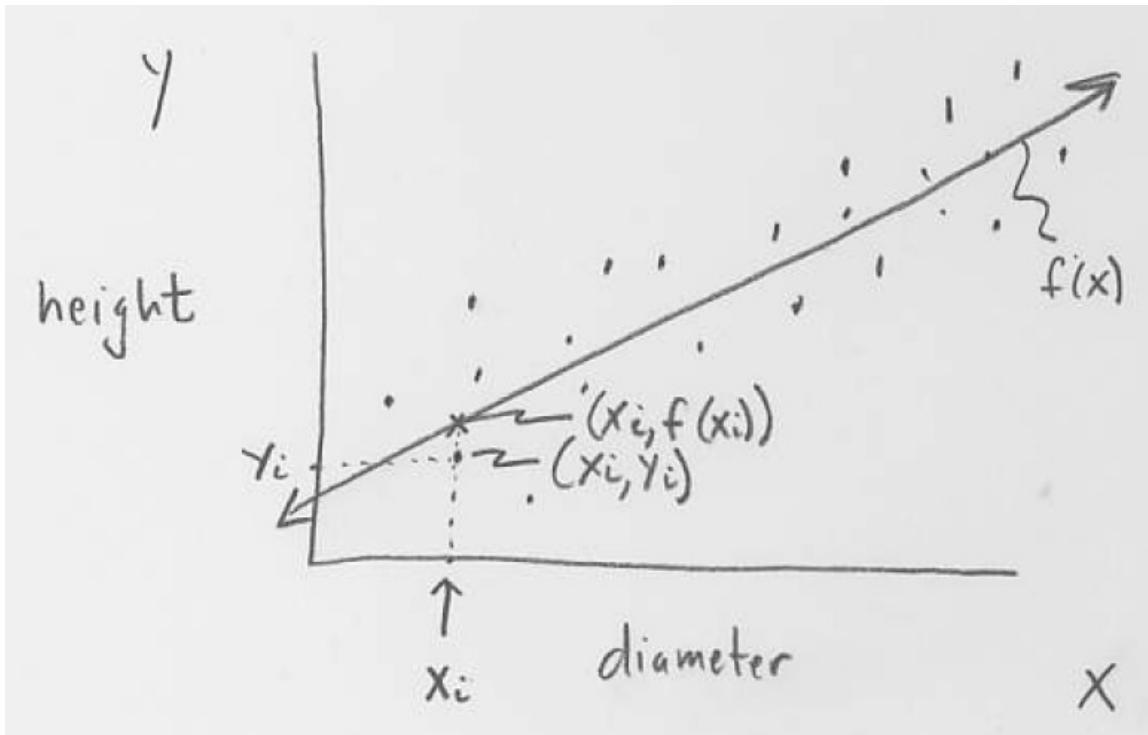
$$\mathbf{B}_{2 \times 3} = [b_{ij}] \quad i = 1, \dots, r, j = 1, \dots, c, r = 2, c = 3$$

**Example 4.3** (An r by c Matrix).

$$\mathbf{A}_{r \times c} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2c} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{ic} \\ \vdots & \ddots & \cdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rj} & \cdots & a_{rc} \end{bmatrix}$$

**Example 4.4** (Simple Linear Regression).

(tree diameter, tree height):  $(x_i, y_i)$ ,  $i = 1, \dots, n$



Data Model Assumptions (just an example):

$$\begin{aligned} Y_i &= \mu_i + \varepsilon_i \quad \text{where } \mu_i = \mu(x_i | \beta_0, \beta_1) = \beta_0 + \beta_1 x_i \\ \varepsilon_i &\stackrel{\text{ind}}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n \end{aligned}$$

or, equivalently,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \varepsilon_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2)$$

Typical matrices for such a model:

$$\mathbf{Y}_{n \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X}_{n \times 2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \boldsymbol{\beta}_{2 \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \boldsymbol{\varepsilon}_{n \times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

(we are on our way to matrix formulation of ANOVA/regression (linear) models:  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ )

**Remark 4.1** (Mean Notation).

- As mentioned previously, we will sometimes use notation like  $\mu(Y | x)$  or  $\mu(Y | x, \beta)$  to denote that the expected value (mean) of  $Y$  is a function of some variable,  $x$ , and parameter vector  $\beta$ .
- Alternatively, we may have written, e.g.,  $\mu_i = E(Y_i | x_i \beta) = \beta_0 + \beta_1 x_i$
- (Recall our remark (3.1) about the mean of a linear function/combination of random variables.)

**Remark 4.2** (Just Matrices and Vectors for the Moment).

- Strictly speaking, at this point, we are discussing matrices/vectors, and, of course, many of our examples are drawn from regression/ANOVA models.
- While we may have some familiarity with regression/ANOVA from previous courses (e.g., STA 270), at the moment, we are simply discussing matrices/vectors without a fuller context of matrices/vectors or their application to regression/ANOVA; we should allow the semester to progress before we become familiar with regression/ANOVA in a matrix context.
- In any case, the examples here may provide insight to regression/ANOVA, later; feel free to think—and ask questions—about regression/ANOVA here, of course!
- While we will look at some of the matrix innards of regression/ANOVA, ultimately, in practice, almost all such details will be held behind the scenes as `R` or `Stan` does most of the work, though this is not entirely true for us in this class; some knowledge of the details may provide a deeper insight and higher level of confidence when using `R` or `Stan` functions for regression/ANOVA.

The following chunk produces an “**X** matrix” (from simulated data) that we may use in regression analysis if we were prepared.

```
> ### Fake tree data (shhh!)
> xdiam<- c(20,15,21,34,28,19,22,24,25,18)
> n<- length(xdiam)
> yht <- 4.5 + 0.8 * xdiam + rnorm(n=n, sd=0.5)
>
> ### We might like to keep our data in a data frame:
> tree.df<- cbind.data.frame(diam=xdiam, height=yht)
>
> ### The X matrix that would be used in regression:
> (X<- model.matrix(height ~ diam, data=tree.df))

  (Intercept) diam
1            1    20
2            1    15
3            1    21
4            1    34
5            1    28
6            1    19
7            1    22
8            1    24
9            1    25
10           1    18
attr(,"assign")
[1] 0 1

> dim(X) ## or size of the X matrix
[1] 10  2
```

## 4.2 Matrix Arithmetic

Matrix/vector operations were implicit, above. Now we make them explicit.

### 4.2.1 Addition/Subtraction

To add/subtract matrices, add/subtract values in the same row/column—easy!

**Example 4.5** (Adding Matrices).

$$\mathbf{A} = \begin{bmatrix} 1 & 4 \\ 2 & 7 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 2 & -3 \\ 4 & 1 \end{bmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 1+2 & 4+(-3) \\ 2+4 & 7+1 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 6 & 8 \end{bmatrix}$$

**Example 4.6** (Adding Matrices (cont'd)).

$$\mathbf{A}_{r \times c} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rc} \end{bmatrix} \quad \mathbf{B}_{r \times c} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1c} \\ b_{21} & b_{22} & \cdots & b_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{rc} \end{bmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1c} + b_{1c} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2c} + b_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} + b_{r1} & a_{r2} + b_{r2} & \cdots & a_{rc} + b_{rc} \end{bmatrix},$$

or

$$[a_{ij}]_{rc} + [b_{ij}]_{rc} = [a_{ij} + b_{ij}]_{rc}$$

**Definition 4.3** (Conformable for Addition).

- *Matrices are said to be conformable for addition if they have the same size.*

- Matrix addition is not defined for matrices of different sizes.

The following Chunk illustrates some matrix functionality in R.

```
> (A<- matrix(c(1,2,4,7),nrow=2, ncol=2)); dim(A)
      [,1] [,2]
[1,]    1    4
[2,]    2    7
[1] 2 2

> (B<- matrix(c(2,4,-3,1),nrow=2, ncol=2)); dim(B)
      [,1] [,2]
[1,]    2   -3
[2,]    4    1
[1] 2 2

> (C<- A + B); dim(C)
      [,1] [,2]
[1,]    3    1
[2,]    6    8
[1] 2 2

> is.matrix(C)
[1] TRUE

> (D<- matrix(c(1,2,3),nrow=3)); dim(D)
      [,1]
[1,]    1
[2,]    2
[3,]    3
[1] 3 1

> A+D
Error in A + D: non-conformable arrays
```

### 4.2.2 Multiply a Matrix by a Scalar

To multiply a matrix by a scalar (single number), multiply each element of the matrix by the scalar—easy!

**Example 4.7** (Scalar Multiplication).

$$a\mathbf{B}_{rc} = a [b_{ij}]_{rc} = [ab_{ij}]_{rc} = \begin{bmatrix} ab_{11} & ab_{12} & \cdots & ab_{1c} \\ ab_{21} & ab_{22} & \cdots & ab_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ ab_{r1} & ab_{r2} & \cdots & ab_{rc} \end{bmatrix}$$

**Example 4.8** (Scalar Multiplication (cont'd)).

$$-2 \begin{bmatrix} 1 & 7 \\ 6 & -3 \end{bmatrix} = \begin{bmatrix} -2(1) & -2(7) \\ -2(6) & -2(-3) \end{bmatrix} = \begin{bmatrix} -2 & -14 \\ -12 & 6 \end{bmatrix}$$

### 4.2.3 Matrix Multiplication

First, some specific examples to illustrate how to multiply matrices.

**Example 4.9** (Matrix Multiplication).

$$\begin{aligned} \begin{bmatrix} 1 & 4 & 7 \\ 6 & 3 & 2 \end{bmatrix}_{2 \times 3} \times \begin{bmatrix} 1 & 2 \\ 6 & 7 \\ 3 & 4 \end{bmatrix}_{3 \times 2} \\ = \begin{bmatrix} 1 \times 1 + 4 \times 6 + 7 \times 3 & 1 \times 2 + 4 \times 7 + 7 \times 4 \\ 6 \times 1 + 3 \times 6 + 2 \times 3 & 6 \times 2 + 3 \times 7 + 2 \times 4 \end{bmatrix}_{2 \times 2} \\ = \begin{bmatrix} 46 & 48 \\ 30 & 41 \end{bmatrix}_{2 \times 2} \end{aligned}$$

The following Chunk illustrates more matrix functionality in R.

```
> A<- matrix(c(1,6,7,-3),nrow=2, ncol=2)
> ## * is for scalar multiplication,
> ## %*% is for matrix multiplication (IMPORTANT!)
> (C<- -2 * A)

      [,1] [,2]
[1,]    -2   -14
[2,]   -12     6

> A<- matrix(c(1,6,4,3,7,2),nrow=2, ncol=3)
> B<- matrix(c(1,6,3,2,7,4),nrow=3,ncol=2)
> (C<- A%*%B)

      [,1] [,2]
[1,]    46   58
[2,]    30   41
```

More generally and abstractly, let

$$\mathbf{C}_{r \times c} = \mathbf{A}_{r \times l} \mathbf{B}_{l \times c}.$$

Then,

$$[c_{ij}]_{r \times c} = \left[ \sum_{k=1}^l a_{ik} b_{kj} \right]_{r \times c}$$

**Definition 4.4** (Conformable for Multiplication).

- Two matrices are said to be **conformable** for multiplication if the left matrix factor in the product (left multiplicand) has the same number of columns ( $l$ ) as the right factor has rows ( $l$ ), else matrix multiplication is not defined.
- Matrix multiplication is otherwise not defined.

```
> A%*%D
      [,1]
[1,]   30
[2,]   18
> D%*%A
Error in D %*% A: non-conformable arguments
```

#### 4.2.4 Matrix Transpose

**Example 4.10** (Matrix Transpose).

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 7 \\ 6 & 3 & 2 \end{bmatrix}_{2 \times 3}$$

$$\mathbf{A}' = \begin{bmatrix} 1 & 6 \\ 4 & 3 \\ 7 & 2 \end{bmatrix}_{3 \times 2}$$

- That is, the element in the  $i$ th row and  $j$ th column of  $\mathbf{A}$  becomes the element in the  $j$ th row and  $i$ th column of the transposed matrix, i.e.,

$$[a_{ij}]'_{r \times c} = [a_{ji}]_{c \times r}.$$

- The transpose operation is often denoted with “prime,”  $\mathbf{A}'$ , or by  $\mathbf{A}^t$  or  $\mathbf{A}^T$  (the latter notation used in [Wak13]).

**Example 4.11** (SLR Continued: Simple Matrix Operations). *Strictly speaking, we now have enough matrix know-how to write down a linear model (e.g., [Wak13, §5.3]). We continue to use the matrices of the SLR Example 4.4. We can write our response vector,  $\mathbf{Y}$ , as a matrix sum of a mean vector—which itself is a matrix product—and error vector (again, we learn regression and linear model details later):*

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{where, again,}$$

$$\mathbf{Y}_{n \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X}_{n \times 2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \boldsymbol{\beta}_{2 \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \boldsymbol{\epsilon}_{n \times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

**Example 4.12** (SLR Continued: Using Transpose and Product). *The matrices  $\mathbf{X}'\mathbf{X}$  and  $\mathbf{X}'\mathbf{Y}$  are always behind the scenes in linear statistical models:*

$$\begin{aligned} \mathbf{X}'\mathbf{X} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \\ &= \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \end{aligned}$$

$$\begin{aligned}\mathbf{X}'\mathbf{Y} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{bmatrix}\end{aligned}$$

The R transpose function, `t()`, is illustrated in the following Chunk.

```
> ### A matrix and its transpose
> (A<- matrix(c(1,6,4,3,7,2), nrow=2, ncol=3))

[,1] [,2] [,3]
[1,]    1    4    7
[2,]    6    3    2

> t(A)

[,1] [,2]
[1,]    1    6
[2,]    4    3
[3,]    7    2

> ### Always a symmetric result.
> t(A) %*% A

[,1] [,2] [,3]
[1,]   37   22   19
[2,]   22   25   34
[3,]   19   34   53

> ### So is this, which is generally different!
> A %*% t(A)

[,1] [,2]
[1,]   66   32
[2,]   32   49
```

### 4.2.5 Special Matrices

**square:**  $r = c$ , e.g.,  $\mathbf{B}_{2 \times 2} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

**symmetric:**  $[a_{ij}] = [a_{ji}]$  (necessarily square, right?!)

$$\text{e.g., } \mathbf{B}_{2 \times 2} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$$

e.g., (SLR cont'd)

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

$\mathbf{XX}'$  is symmetric, too, but we do not use it in this class.

**identity:** All ones on diagonal, e.g.,

$$\mathbf{I}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Used often. Note  $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$  for any conformable  $\mathbf{A}$  and  $\mathbf{I}$ . The matrix version of number 1.

**diagonal:** All zeros off main (upper left to lower right) diagonal

e.g.,

$$\mathbf{A}_{rr} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & a_{rr} \end{bmatrix}_{r \times r}$$

e.g.,

$$\begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}_{n \times n}$$

e.g., often in regression/ANOVA

$$\sigma^2 \mathbf{I}$$

**unity**: vector:

$$\mathbf{1}_{n \times 1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1}$$

matrix:

$$\mathbf{J}_{n \times n} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}_{n \times n}$$

Note  $\mathbf{J} = \mathbf{1}\mathbf{1}'$  and  $n = \mathbf{1}'_{n \times 1} \mathbf{1}_{n \times 1}$

**zero**: vector:

$$\mathbf{0}_{n \times 1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n \times 1}$$

matrix:

$$\mathbf{0}_{n \times n} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}_{n \times n}$$

Note  $\mathbf{0} = \mathbf{0A}$

**unit vector in ith coordinate direction or standard unit vector**: One

in  $i$ th element zeros elsewhere,

$$\mathbf{e}_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n \times 1} .$$

The  $i$ th column of  $\mathbf{I}_{n \times n}$ .

The following Chunk illustrates identity matrices in R.

```
> ### Quick way to create identity matrices
> (I2<- diag(2))

 [,1] [,2]
[1,]    1    0
[2,]    0    1

> (I3<- diag(3))

 [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1

> ### Identity matrices in action (or lack thereof)
> I2%*%A

 [,1] [,2] [,3]
[1,]    1    4    7
[2,]    6    3    2

> A%*%I3

 [,1] [,2] [,3]
[1,]    1    4    7
[2,]    6    3    2
```

#### 4.2.6 Linear Dependence & Rank

**Definition 4.5** (Linear Combination (of vectors):). Let  $\mathbf{a}_i$ ,  $i = 1, \dots, n$ , be column matrices (i.e., vectors) of the same size (same number of rows). Let  $c_i$ ,  $i = 1, \dots, n$  be some scalars (numbers not matrices). Then

$$\sum_{i=1}^n c_i \mathbf{a}_i = c_1 \mathbf{a}_1 + \cdots + c_n \mathbf{a}_n$$

is a **linear combination** of the  $\mathbf{a}_i$  vectors with **coefficients**  $c_i$ .

**Example 4.13** (Linear Combination).

$$\mathbf{a}_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \quad \mathbf{a}_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \quad \mathbf{a}_3 = \begin{bmatrix} 8 \\ 5 \end{bmatrix},$$

$c_1 = c_2 = c_3 = 1$ . Then,

$$\begin{aligned} \sum_{i=1}^n c_i \mathbf{a}_i &= (1) \begin{bmatrix} 2 \\ 3 \end{bmatrix} + (1) \begin{bmatrix} 3 \\ 1 \end{bmatrix} + (1) \begin{bmatrix} 8 \\ 5 \end{bmatrix} \\ &= \begin{bmatrix} 13 \\ 9 \end{bmatrix} \end{aligned}$$

**Definition 4.6** (Linear Dependence (of vectors)). Vectors  $\mathbf{a}_i$ ,  $i = 1, \dots, n$ , are said to be **linearly dependent** if there exists numbers  $c_i$ ,  $i = 1, \dots, n$ , not all zero, such that

$$c_1 \mathbf{a}_1 + \cdots + c_n \mathbf{a}_n = \mathbf{0}.$$

Else, the  $a_i$  are said to be **linearly independent** (different than independence of rvs).

**Example 4.14** (Linear Dependence). *Continuing the above Example 4.13, we see  $\mathbf{a}_3 = \mathbf{a}_1 + 2\mathbf{a}_2$ , i.e.,  $\mathbf{a}_1 + 2\mathbf{a}_2 + (-1)\mathbf{a}_3 = \mathbf{0}$ , i.e.,*

$$(1) \begin{bmatrix} 2 \\ 3 \end{bmatrix} + 2 \begin{bmatrix} 3 \\ 1 \end{bmatrix} + (-1) \begin{bmatrix} 8 \\ 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

*Thus, the  $\mathbf{a}_i$  are said to be linearly dependent. Given a subset of (two in this case) the vectors, we can reproduce the remaining (one in this case) vectors.*

**Example 4.15** (Multiple Linear Regression (MLR):). *Using the set-up in the SLR examples above (4.4, 4.11, 4.12), suppose your colleague wishes to include a second “x” variable in the following manner:*

$$x_{i2} = 2x_{i1},$$

*where we now include extra subscripts to distinguish our two covariates. (Say the  $x_{i1}$  are tree diameters, previously denoted just as  $x_i$ , so that, here, the  $x_{i2}$  are just twice the diameter values.) In this case, the  $\mathbf{X}$  matrix for (multiple) linear regression would be*

$$\begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$$

*But, because of the way your colleague created the  $x_{2i}$  values, we know*

$$(0) \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + (-2) \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} + (1) \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} = \mathbf{0}.$$

*This sort of dependence will lead to problems in regression. (Admittedly,*

*it seems a bit contrived, but you'd be surprised what people do to their data!)*

**Definition 4.7** (Rank). *The (maximum) number of linearly independent columns (or rows) of matrix. (Thus,  $\text{rank}(\mathbf{A}_{r \times c}) \leq \min(r, c)$ .) In other words, rank is the maximum number of columns (rows) that can be selected before one of the selected vectors can be reproduced by a linear combination of the other selected vectors.*

**Example 4.16** (Rank). *In Example 4.13 we have*

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix} | & | & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \\ | & | & | \end{bmatrix} \\ &= \begin{bmatrix} 2 & 3 & 8 \\ 3 & 1 & 5 \end{bmatrix}\end{aligned}$$

$\text{rank}(\mathbf{A}) = 2$  Why?

For the strange MLR Example 4.15 we have  $\text{rank}(\mathbf{X}) = 2$ , assuming  $n \geq 2$  distinct diameters (usually  $n \gg 2$ ).

**Definition 4.8** (Inverse (of square matrix)). *Let  $\mathbf{A}$  be square. The inverse of  $\mathbf{A}$ , if it exists, is denoted  $\mathbf{A}^{-1}$ , and is such that*

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A}.$$

e.g., Let  $\mathbf{A} = [6]$ . Then  $\mathbf{A}^{-1} = [1/6]$  and  $[1/6][6] = [1] = [6][1/6]$ .

e.g.,

$$\begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 2 & -1 \\ -3 & 2 \end{bmatrix}$$

because

$$\begin{bmatrix} 2 & -1 \\ -3 & 2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -3 & 2 \end{bmatrix}.$$

**NOTE:** Perhaps obviously,  $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$  (inverse of inverse is the original matrix).

**NOTE:**  $(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}$  and is often denoted  $\mathbf{A}^{-T}$  while using  $'$  doesn't seem fashionable.

The next chunk illustrates the computation of a matrix inverse in R.

```
> ### A matrix and its inverse.
> set.seed(8675309)
> (B<- matrix(round(10*rnorm(16)), nrow=4, ncol=4))

      [,1] [,2] [,3] [,4]
[1,]   -10    11     6     2
[2,]     7    10     9    -7
[3,]    -6     0   -15   -10
[4,]    20     7    10    20

> (Binv<- solve(B))

      [,1]          [,2]          [,3]          [,4]
[1,] -0.0423590131  0.03033736  0.0167421954  0.02322508
[2,]  0.0478348439  0.02517623  0.0553877140  0.03172205
[3,] -0.0002014099  0.02094663 -0.0739174220 -0.02960725
[4,]  0.0257175227 -0.04962236  0.0008308157  0.03047583

> round(Binv%*%B, 4)
```

```
[,1] [,2] [,3] [,4]
[1,] 1 0 0 0
[2,] 0 1 0 0
[3,] 0 0 1 0
[4,] 0 0 0 1
```

```
> round(B %*% Binv, 4)
```

```
[,1] [,2] [,3] [,4]
[1,] 1 0 0 0
[2,] 0 1 0 0
[3,] 0 0 1 0
[4,] 0 0 0 1
```

## A Few More Things

- If  $\mathbf{A}^{-1}$  exists, then  $\mathbf{A}$  is said to be **non-singular**, else it's **singular**. (Square matrix implied here.)
- If  $\text{rank}(\mathbf{A}_{r \times c}) = \min(r, c)$ , then  $\mathbf{A}$  is said to be **full rank**, else, if  $\text{rank}(\mathbf{A}_{r \times c}) < \min(r, c)$ , then it is not full rank, and we say  $\mathbf{A}$  is **rank deficient**.
- If  $\mathbf{A}$  is square, then  $\mathbf{A}$  being full rank is equivalent to existence of  $\mathbf{A}^{-1}$ .
- $\text{rank}(AB) = \min(\text{rank}(A), \text{rank}(B))$

**Example 4.17** (Regression with 2 or More  $X$  Variables (MLR)).

***Modeling Assumptions:***

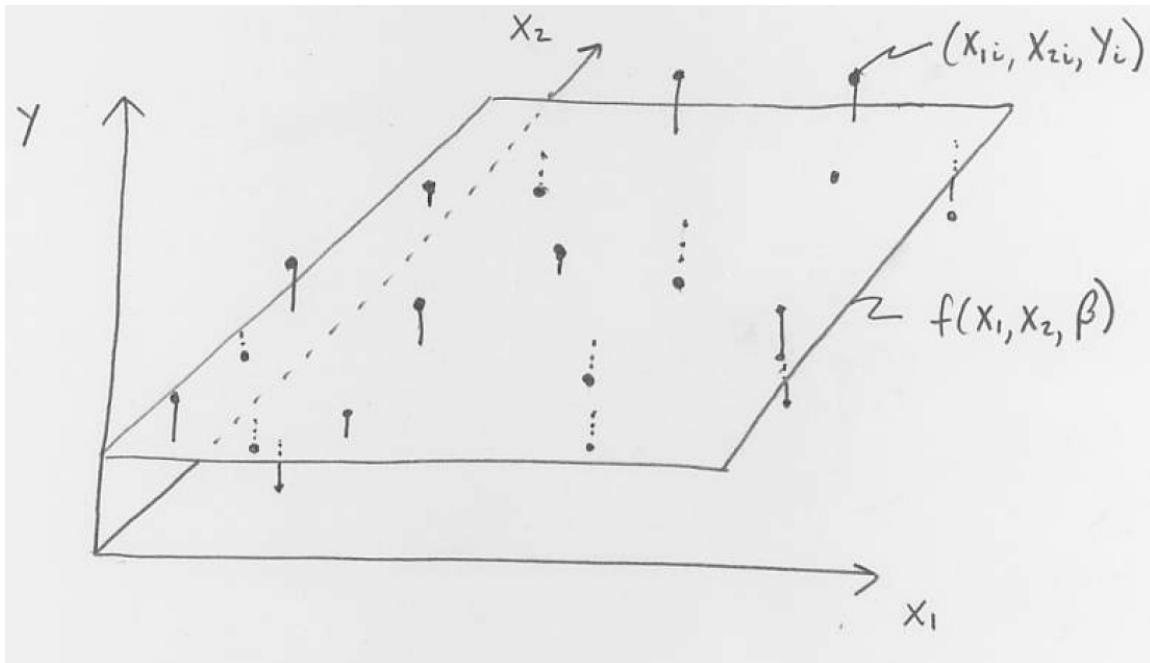
$$Y_i = \mu_i + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2) \quad i = 1, \dots, n$$

where

$$\mu_i = \overbrace{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)}}^{\mu(Y|X_{i1}, \dots, X_{i(p-1)})}$$

i.e.,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)} + \varepsilon_i.$$



(Can you find notational inconsistency in above figure?)

**Major Goal:** not surprisingly, estimate the mean(s),  $\mu_i$ , which, in this case, means estimate

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}.$$

We will also estimate  $\sigma^2$ . On the way to estimating  $\boldsymbol{\beta}$ , we come to the **normal equations** (unrelated to the normal distribution):

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y} \quad \text{details omitted.}$$

If  $(\mathbf{X}'\mathbf{X})^{-1}$  exists, we left-multiply both sides of the above to get

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

$\hat{\boldsymbol{\beta}}$  is the **estimator** or **estimate** of the unknown parameter  $\boldsymbol{\beta}$ . (We wait to estimate  $\sigma^2$ . Remember, we're still just doing matrix manipulations at this point, though some of you might be "getting" the regression stuff a bit.)

### Example 4.18 (Rank Deficiency and Redundancy in Regression).

- In regression/ANOVA, the number of rows,  $n$ , of  $\mathbf{X}$  ( $n$  is number of observations) is typically greater than the number of columns/mean parameters ( $n > p$ ).
- Thus, if  $\mathbf{X}$  is full rank,  $\text{rank}(\mathbf{X}) = \min(n, p) = p$  and, by a previously stated result,  $\text{rank}(\mathbf{X}'\mathbf{X}) = \min(\text{rank}(\mathbf{X}'), \text{rank}(\mathbf{X})) = \min(\min(n, p), \min(n, p)) = \min(n, p) = p$ .
- That is  $\mathbf{X}'\mathbf{X}$  is full-rank ( $= p$ ), and is non-singular, and  $(\mathbf{X}'\mathbf{X})^{-1}$  exists, and we may solve for  $\boldsymbol{\beta}$ .
- Else, if  $\text{rank}(\mathbf{X}) < p$ , then  $\mathbf{X}'\mathbf{X}$  is singular and the inverse does not exist. In this case, we must resort to "rank deficient" methods, which we do not cover.
- We will encounter briefly rank deficient  $\mathbf{X}$  when discussing ANOVA, but will introduce fairly straightforward ways to get to an equivalent (in some sense) full rank  $\mathbf{X}$  so that  $\mathbf{X}'\mathbf{X}$  is invertible to get solutions to our linear models.
- If  $\mathbf{X}$  is not full-rank, that means we have linear dependence among the columns of  $\mathbf{X}$ ; we can reproduce at least one column from a linear combination of the others.

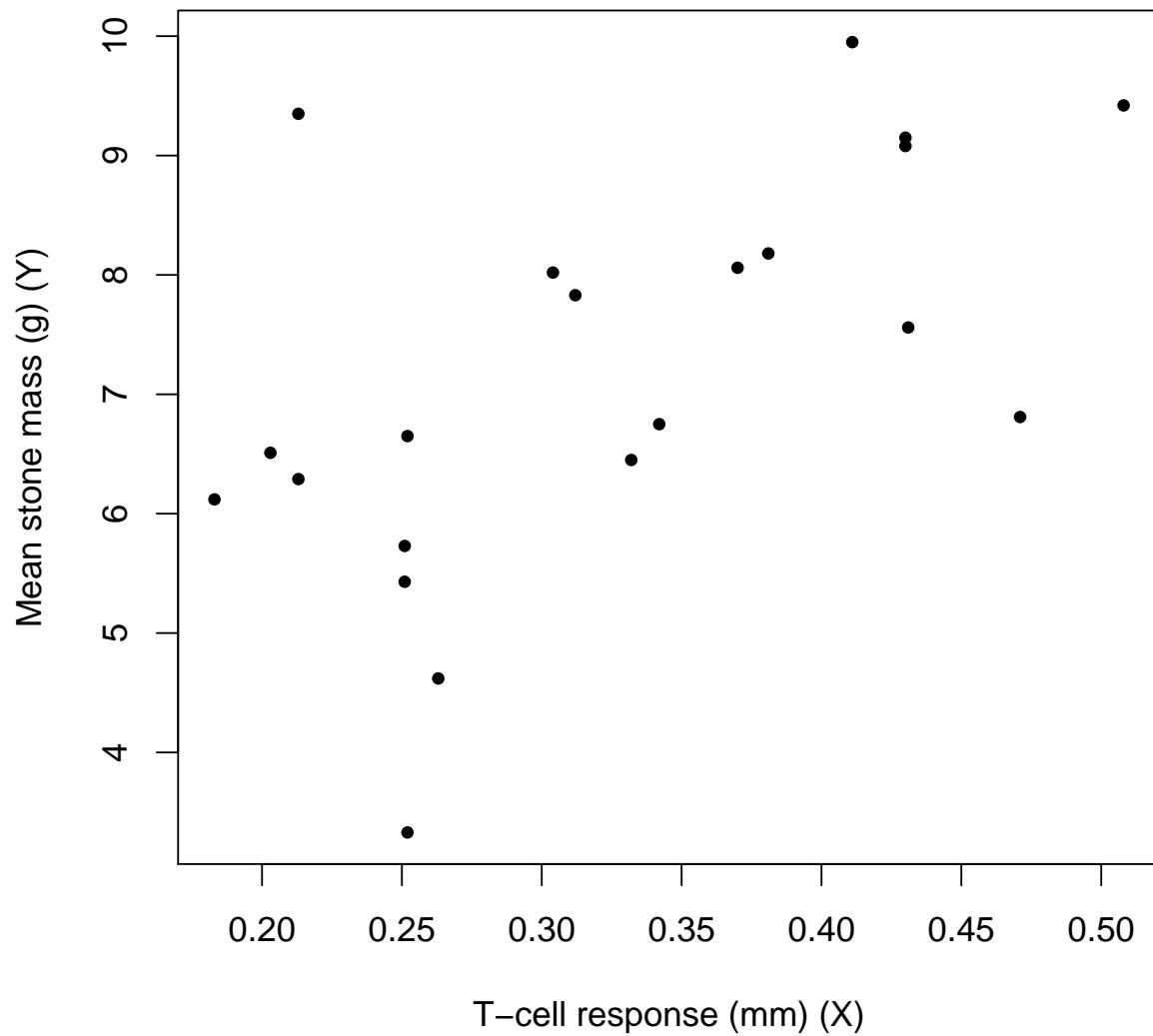
- In this sense, we have redundant information amongst our  $X$  variables and may be able to omit one or more. See Example 4.15.
- Related to multicollinearity and non-identifiability and non-estimability of mean parameters and variable selection. Perhaps more later.

See [KNNL05, Sec. 5.7] for more basic matrix results.

The following Chunk illustrates even more matrix functionality in R in the context of regression. Again, we haven't arrived at regression yet. We are simply illustrating some of the matrix functionality presented above. Of course, it's meant to show also that matrices are useful for regression/ANOVA computations.

```
> ### Data
> ex0727.df<- Sleuth3::ex0727
>
> ### Plot:
> par(cex=1.2)
> plot(Mass ~ Tcell, data=ex0727.df,
+       pch=20,
+       main="Weight-lifting and Bird Health",
+       xlab="T-cell response (mm) (X)",
+       ylab="Mean stone mass (g) (Y)")
```

## Weight-lifting and Bird Health



```
> X<- model.matrix(Mass ~ Tcell + I(Tcell^2), data= ex0727.df)
> head(X, n=3); tail(X,n=3)

(Intercept) Tcell I(Tcell^2)
1           1 0.252  0.063504
2           1 0.263  0.069169
3           1 0.251  0.063001
(Intercept) Tcell I(Tcell^2)
19          1 0.213  0.045369
```

```
20      1 0.508  0.258064
21      1 0.411  0.168921

> dim(X)

[1] 21  3

> n<- dim(X)[1]
> Y<- ex0727.df$Mass
>
> (beta_hat<- solve(t(X) %*% X) %*% t(X) %*% Y)

[,1]
(Intercept) 5.766853
Tcell       -1.801225
I(Tcell^2)  17.749740

> coefficients(lm(Mass ~ Tcell + I(Tcell^2), data = ex0727.df))

(Intercept)      Tcell    I(Tcell^2)
5.766853     -1.801225   17.749740
```

## 4.3 Combining Things: Random Vectors and Matrices

Not surprisingly, a random matrix is a matrix whose elements are random variables. We focus on vectors.

Just like random (scalar) variables, random vectors are associated with probability models (pdf, pmf), but we will minimize discussion of these joint models, for the moment, and focus instead on easy-to-understand summarizing properties of these models. We will cover joint distributions as well as marginal distributions and conditional distributions in a subsequent chapter.

### 4.3.1 Expectation of a Random Vector/Matrix

If

$$\mathbf{Y}_{n \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

is an rv, then

$$E(\mathbf{Y}) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix}$$

In words: *expected value of a random vector is the vector of expected values.*

**Example 4.19** (Expected Value of Normal RV). If we assume  $Y_i \stackrel{\text{ind}}{\sim} N(\mu, \sigma^2)$ , then

$$E(\mathbf{Y}) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [\mu]$$

This result has nothing to do with independence.

### 4.3.2 Variance(-Covariance) Matrix

$$\text{Var}(\mathbf{Y}) = \begin{bmatrix} \text{Var}(Y_1) & \text{Cov}(Y_1, Y_2) & \cdots & \text{Cov}(Y_1, Y_n) \\ \text{Cov}(Y_2, Y_1) & \text{Var}(Y_2) & \cdots & \text{Cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_n, Y_1) & \text{Cov}(Y_n, Y_2) & \cdots & \text{Var}(Y_n) \end{bmatrix}$$

which can be written as

$$\begin{aligned}
 [\text{Cov}(Y_i, Y_j)] &= [\mathbb{E}((Y_i - \mathbb{E}(Y_i))(Y_j - \mathbb{E}(Y_j)))] \\
 &= \mathbb{E} \left( \begin{bmatrix} Y_1 - \mathbb{E}(Y_1) \\ Y_2 - \mathbb{E}(Y_2) \\ \vdots \\ Y_n - \mathbb{E}(Y_n) \end{bmatrix} \begin{bmatrix} Y_1 - \mathbb{E}(Y_1) & Y_2 - \mathbb{E}(Y_2) & \cdots & Y_n - \mathbb{E}(Y_n) \end{bmatrix} \right) \\
 &= \mathbb{E}((\mathbf{Y} - \mathbb{E}(\mathbf{Y}))(\mathbf{Y} - \mathbb{E}(\mathbf{Y}))')
 \end{aligned}$$

In words, *the variance of a random vector is the matrix of variances and covariances computed from the elements of the random vector.*

#### 4.3.3 Linearity of Expectation Operator (just as in scalar case)

$$\begin{aligned}
 \mathbb{E}(\mathbf{a} + \mathbf{B}\mathbf{Y}) &= \mathbb{E}(\mathbf{a}) + \mathbb{E}(\mathbf{B}\mathbf{Y}) \\
 &= \mathbf{a} + \mathbf{B}\mathbb{E}(\mathbf{Y})
 \end{aligned}$$

where  $\mathbf{a}$  is a vector of constants,  $\mathbf{B}$  is a matrix of constants, and  $\mathbf{Y}$  is an rv. We gave an (almost) analogous result in terms of scalars in a previous remark (3.1).

#### 4.3.4 Variance(-Covariance) of $\mathbf{a} + \mathbf{B}\mathbf{Y}$

$$\begin{aligned}
 \text{Var}(\mathbf{a} + \mathbf{B}\mathbf{Y}) &= \text{Var}(\mathbf{a}) + \text{Var}(\mathbf{B}\mathbf{Y}) \\
 &= 0 + \mathbf{B}\text{Var}(\mathbf{Y})\mathbf{B}' \\
 &= \mathbf{B}\text{Var}(\mathbf{Y})\mathbf{B}',
 \end{aligned}$$

which reduces to result given back in §3.3.2 when  $\mathbf{a}$ ,  $\mathbf{Y}$  and  $\mathbf{B}$  are scalars.

#### 4.3.5 Distribution of Linear Function of Normal RV

This next result, along with the results in the previous subsections, 4.3.3 and 4.3.4, will be used repeatedly throughout the course, though the details will often be behind the scenes, and the notation may change with context. The scalar analog was given in §3.6.

If  $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  (see [KNNL05, p.197] for pdf of such a vector), then  
 $(\mathbf{a} + \mathbf{B}\mathbf{Y}) \sim N(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$

**Example 4.20** (Introductory Statistics Model in Matrix Form).

Assume

$$Y_i \stackrel{ind}{\sim} N(\mu, \sigma^2) \quad i = 1, \dots, n$$

Equivalently,

$$Y_i = \mu + \varepsilon_i \quad \varepsilon_i \stackrel{ind}{\sim} N(0, \sigma^2) \quad i = 1, \dots, n.$$

(We have a result for this equivalence, right?! Note, by the way,  $\varepsilon_i = Y_i - \mu = Y_i - E(Y_i)$ .)

“Stacking things” and using matrix/vector multiplication/addition, we can write

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \boldsymbol{\beta} = [\mu], \quad \boldsymbol{\epsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

That is,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [\mu] + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Using our previously presented results, we know

•

$$\begin{aligned} E(\mathbf{Y}) &= E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = E(\mathbf{X}\boldsymbol{\beta}) + E(\boldsymbol{\epsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{0} = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix}, \end{aligned}$$

•

$$\begin{aligned}
 Var(\boldsymbol{\epsilon}) &= [Cov(\varepsilon_i, \varepsilon_j)] \\
 &= \begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1, \varepsilon_2) & \cdots & Cov(\varepsilon_1, \varepsilon_n) \\ Cov(\varepsilon_2, \varepsilon_1) & Var(\varepsilon_2) & \cdots & Cov(\varepsilon_2, \varepsilon_n) \\ \vdots & \ddots & \ddots & \vdots \\ Cov(\varepsilon_n, \varepsilon_1) & Cov(\varepsilon_n, \varepsilon_2) & \cdots & Var(\varepsilon_n) \end{bmatrix} \\
 &= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n
 \end{aligned}$$

•

$$\begin{aligned}
 Var(\mathbf{Y}) &= Var(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
 &= Var(\mathbf{X}\boldsymbol{\beta}) + Var(\boldsymbol{\epsilon}) \\
 &= 0 + Var(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n
 \end{aligned}$$

So, we could denote our model as

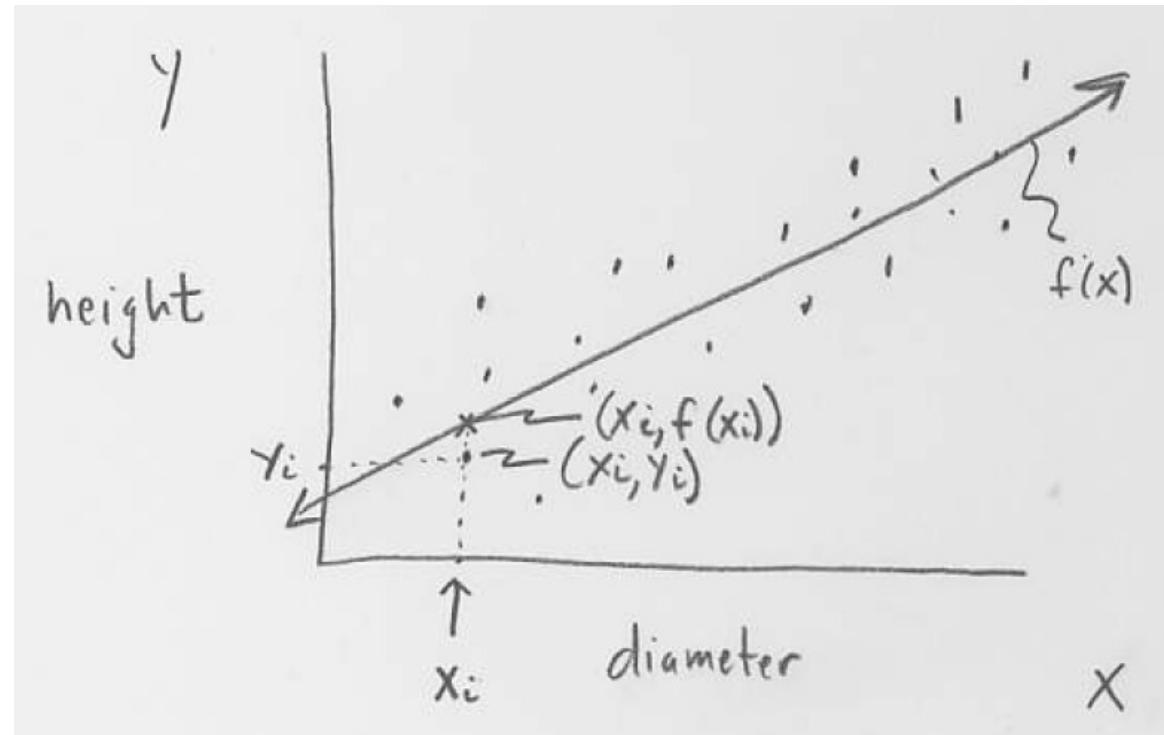
$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n),$$

or as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

We exercise our new-found statistical and matrix skills in a very similar manner for the (hopefully) familiar simple linear regression and 1-way ANOVA and models, below.

**Example 4.21** (SLR Model in Matrix Form).  
*(tree diameter, tree height):  $(X_i, Y_i)$ ,  $i = 1, \dots, n$*



*Model Assumptions (just an example):*

$$\begin{aligned} Y_i &= \mu_i + \varepsilon_i \quad \text{where } \mu_i = \mu(Y | x_i) = \beta_0 + \beta_1 x_i \\ \varepsilon_i &\stackrel{\text{ind}}{\sim} N(0, \sigma^2) \end{aligned}$$

*or, equivalently,*

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \varepsilon_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2),$$

*or*

$$Y_i \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2).$$

*As in the previous example, we can write our model as*

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

*where (now a bit differently),*

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

That is,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

And,

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix},$$

$$Var(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I},$$

$$Var(\mathbf{Y}) = \sigma^2 \mathbf{I}.$$

**Example 4.22** (1-Way ANOVA Model in Matrix Form). Write a 1-way ANOVA model, with, say,  $I = 3$  levels, in the form of a matrix linear

model. (*to be done in class*)

If you haven't gathered that many (regression and ANOVA) models have effectively the same form (and analysis methods (to be discussed)), we will drive this home in subsequent chapters of our notes. After all, this current chapter was, strictly speaking, only meant to introduce matrices, vectors and a few fundamental concepts.

# Lecture 5

## Linear Models I: Introduction

### Contents

---

|            |   |            |
|------------|---|------------|
| <b>5.1</b> | <b>Motivating Data Set . . . . .</b>                              | <b>151</b> |
| <b>5.2</b> | <b>Distributions: Joint, Marginal &amp; Conditional . . . . .</b> | <b>153</b> |
| <b>5.3</b> | <b>Model Specification . . . . .</b>                              | <b>160</b> |
| 5.3.1      | It's a Conditional Mean Model Specification . . . . .             | 161        |
| 5.3.2      | Covariates Observed Without Error . . . . .                       | 163        |
| <b>5.4</b> | <b>A Justification of Linear Modeling . . . . .</b>               | <b>163</b> |
| <b>5.5</b> | <b>Parameter Interpretation . . . . .</b>                         | <b>165</b> |
| 5.5.1      | Conditional Mean Model vs. Marginal Mean model . . . . .          | 165        |
| 5.5.2      | Extrapolation, Meaningful Parameters & Reparameterization . . .   | 166        |
| 5.5.3      | Typical “Additive Change” Parameter Interpretation . . . . .      | 168        |
| 5.5.4      | Data Transformations . . . . .                                    | 173        |

---

***Main Objectives:***

- Introduction to joint, marginal and conditional distributions
- Multiplication rule for joint distributions
- Independence of random variables
- Regression function is the conditional mean

- Regression is typically conditional on covariates. This implies that the distribution of covariates does not depend on regression function parameters (or error variance).
  - Covariates are assumed to have no error
  - Modeling the regression function as linear may be viewed as a first order Taylor approximation to the unknown regression function
  - Beware of extrapolating beyond observed data
  - Beware of regression function parameter interpretation
- 
- $\mathcal{O}$

***Additional Reading:***

- [Wak13, §5.1 - 5.5] and selected other references to [Wak13].

---

 $\mathcal{R}$ 

## 5.1 Motivating Data Set

- We use the `Prostate` (cancer) data set in the `lasso2` package.
- Observational study.
- You will find more information via `help(Prostate)` (after submitting `library(lasso2)`).
- `lpsa` ( $\log(\text{PSA})$ , prostate specific antigen) is the response,  $y$ , with  $k = 8$  potential covariates
- Used throughout [Wak13, Chap. 5].
- We'll see other data sets, too!

```
> library(lasso2)

R Package to solve regression problems while imposing
an L1 constraint on the parameters. Based on S-plus Release 2.1
Copyright (C) 1998, 1999
Justin Lokhorst <jlokhors@stats.adelaide.edu.au>
Berwin A. Turlach <bturlach@stats.adelaide.edu.au>
Bill Venables <wvenable@stats.adelaide.edu.au>
Copyright (C) 2002
Martin Maechler <maechler@stat.math.ethz.ch>
```

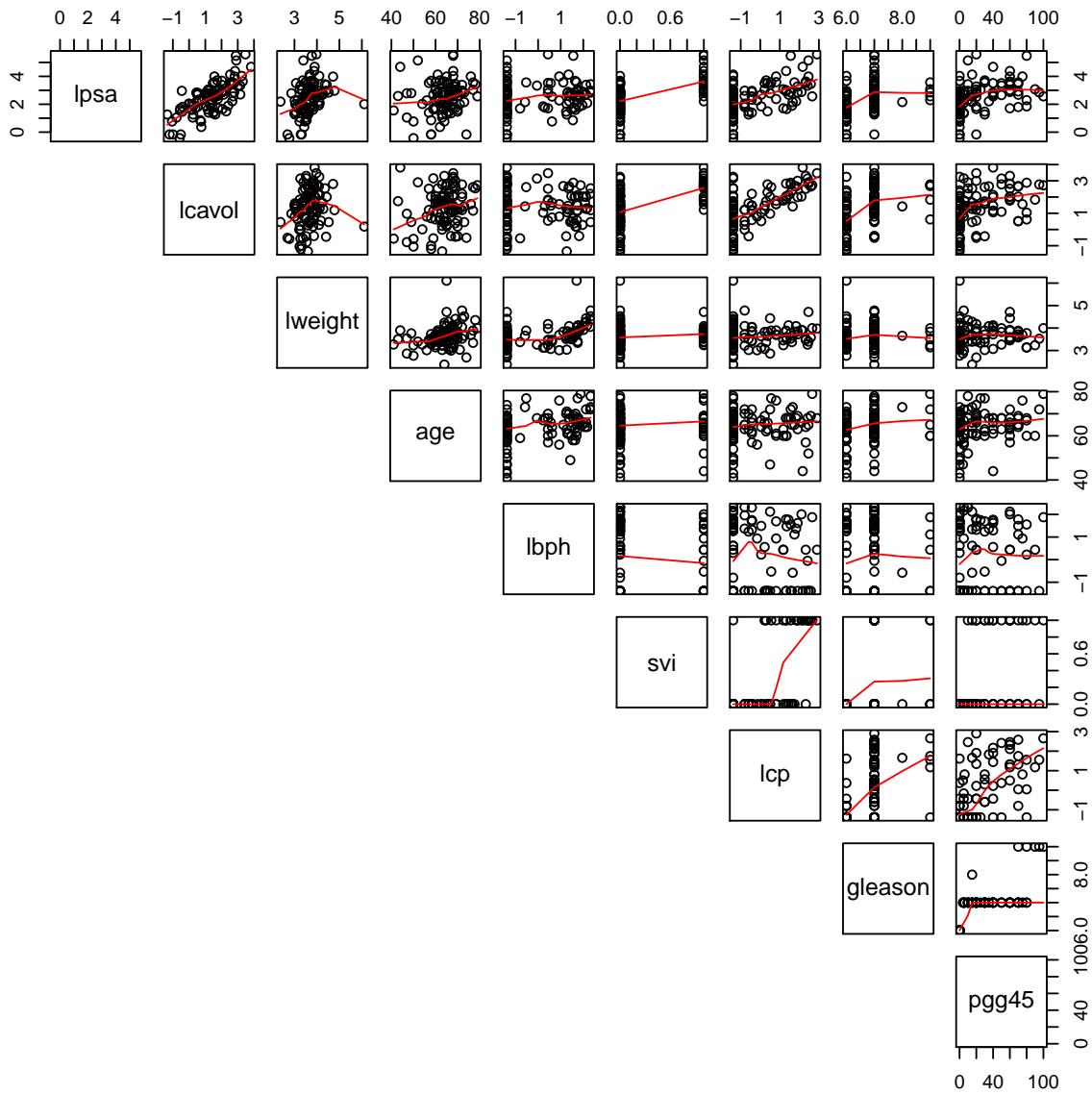
```
> data(Prostate)
> dim(Prostate)

[1] 97 9

> names(Prostate)

[1] "lcavol"   "lweight"  "age"       "lbph"      "svi"
[6] "lcp"       "gleason"  "pgg45"     "lpsa"
```

```
> pairs(lpsa ~ . ,
+        upper.panel = panel.smooth,
+        lower.panel = NULL,
+        data= Prostate)
```



## 5.2 Distributions: Joint, Marginal & Conditional

A joint probability distribution describes how two or more random variables/vectors vary together (“jointly”), and, if known, can be used to answer questions regarding the probability of subsets of values of  $X$  and  $Y$ , e.g.,  $P(X \leq x, Y \leq y)$  for scalar variables.

As for scalar random variables of §3.2, we have analogous definitions of

(joint) probability distributions (or related functions) of random vectors.

**Definition 5.1** (Joint cdf).

- The joint **cumulative distribution function** (cdf) of a random vector  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$  is given by

$$F(y_1, \dots, y_n) = P(Y_1 \leq y_1, \dots, Y_n \leq y_n),$$

where  $P(Y_1 \leq y_1, \dots, Y_n \leq y_n)$  is interpreted as the “probability” of the random variables  $Y_i$  being less than or equal to some numbers  $y_i$ , simultaneously (“jointly”),  $i = 1, \dots, n$ .

- Note the lowercase  $y_i$  represent fixed (non-random) values that must be specified in order to get a value for the function  $F$ , just like a typical mathematical function.
- As we mentioned in §3.2, this definition seems to imply that the (joint) cdf,  $F$ , is defined as a function of some sort of probability function,  $P$ , as if  $P$  exists first.
- In practice, we typically specify the cdf,  $F$ , (or, more likely, the pmf/pdf, below), which induces a corresponding probability function  $P$ , which we do not discuss. Again, we avoid a formal definition of probability.

**Definition 5.2** (Joint pdf).

- If there exists a function  $f(y_1, \dots, y_n)$  such that

$$F(y_1, \dots, y_n) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_n} f(t_1, \dots, t_n) dt_1 \cdots dt_n$$

is a cdf, then  $f(\mathbf{y})$  is called a (joint) **probability density function (pdf)**.

- Our definition of joint **pdf** implies  $Y_i$  is a continuous random variable.
- The support of  $\mathbf{y}$  is implicitly incorporated into  $f(\mathbf{y})$ .
- The analogous definition of the **joint pmf** (probability mass function) for vector of discrete random variables replaces integrals by sums in the above definition.
- In practice, we **typically** specify mathematical functions for the **joint pdf or joint pmf** and do not work nearly as much with cdfs.
- We might also specify a functions corresponding to the joint distribution of a collection of continuous and discrete random variable (not much in this class).

**Example 5.1** (Multivariate Normal pdf). *If  $Y_i$  are independently normal with mean,  $\mu_i$ , and standard deviation,  $\sigma_i$ ,  $i = 1, \dots, n$ , then the (joint) pdf of  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$  is*

$$f(y_1, \dots, y_n) = (2\pi)^{-n/2} \prod_i (\sigma_i) \exp \left( -\frac{1}{2} \sum_i \left( \frac{y_i - \mu_i}{\sigma_i} \right)^2 \right),$$

which we can write in matrix form as

$$f(\mathbf{y}) = (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp \left( -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right),$$

where  $\mathbf{y} = (y_1, \dots, y_n)^T$ ,  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ ,  $\mathbf{V}$  is the  $n \times n$  diagonal matrix

with  $i$ th diagonal element  $\sigma_i^2$ , and  $|\mathbf{V}|$  denotes the determinant of (square) matrix  $\mathbf{V}$ .

- In a **traditional linear model**,  $\mu_i \equiv E(Y_i | \mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i^T \boldsymbol{\beta}$ , where  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$  is a  $p$ -vector of observed covariates corresponding to observed response,  $y_i$ ,  $\boldsymbol{\beta}$  is a  $p$ -vector of mean parameters and  $\sigma_i^2 = \sigma^2$  (constant variance).

```
> ## E.g., cdf computation for vector of n iid N(0,1) rvs
> n<- 30
> ## F(y_1 <= 2, ..., y_{\{n\}} <= 2)
> prod(pnorm(rep(2,n)))

[1] 0.5013819

> ## Or
> library(mvtnorm)
> pmvnorm(lower=-Inf, upper=rep(2,n),
+           mean=rep(0,n), sigma=diag(n))

[1] 0.5013819
attr(,"error")
[1] 0
attr(,"msg")
[1] "Normal Completion"

> detach(package:mvtnorm)
>
> ## Or
> mvtnorm::pmvnorm(lower=-Inf, upper=rep(2,n),
+                   mean=rep(0,n), sigma=diag(n))
```

```
[1] 0.5013819
attr(,"error")
[1] 0
attr(,"msg")
[1] "Normal Completion"
```

**Definition 5.3** (Square Bracket Notation for pdf/pmf).

- Instead of using alphabetical notation, such as  $f(\mathbf{y})$ , we will often use  $[\mathbf{y}]$  to denote the pdf/pmf of random vector  $\mathbf{Y}$ .
- e.g.,  $[\mathbf{x}, \mathbf{y}]$  denotes the joint pdf/pmf of two random vectors  $\mathbf{X}$  and  $\mathbf{Y}$ .
- [Wak13] does not use square-bracket notation.
- Note that we (not [Wak13]) will also use  $\mathbf{X}$  to denote a matrix of observed covariate values. This should cause little confusion as we will typically consider covariates to be fixed in this class (more below).

**Definition 5.4** (Marginal pdf/pmf/distribution).

- If  $[\mathbf{x}, \mathbf{y}]$  is the (joint) pdf (distribution) of two (continuous) random vectors,  $\mathbf{X}$  and  $\mathbf{Y}$ , then the marginal pdfs (distributions) of  $\mathbf{X}$  and  $\mathbf{Y}$  are

$$[\mathbf{x}] = \int [\mathbf{x}, \mathbf{y}] d\mathbf{y},$$

and

$$[\mathbf{y}] = \int [\mathbf{x}, \mathbf{y}] d\mathbf{x}.$$

- Multiple integration implied.
- For discrete random vectors (pmfs), sums replace integrals.

**Definition 5.5** (Conditional pdf/pmf/distribution).

- If  $[\mathbf{x}, \mathbf{y}]$  is the (joint) pdf (distributions) of two random vectors,  $\mathbf{X}$  and  $\mathbf{Y}$ , then the (joint!) conditional pdfs (distributions) are

$$[\mathbf{x} | \mathbf{y}] = \frac{[\mathbf{x}, \mathbf{y}]}{[\mathbf{y}]}$$

and

$$[\mathbf{y} | \mathbf{x}] = \frac{[\mathbf{x}, \mathbf{y}]}{[\mathbf{x}]}.$$

- A **conditional distribution** describes the distribution of one random vector for a given (“|” **conditional on**) fixed value of the other variable/vector
- We denote the associated **conditional random variables** as  $\mathbf{X} | \mathbf{y}$  and  $\mathbf{Y} | \mathbf{x}$ , respectively.

- Joint, marginal and conditional distributions are distributions like any others with corresponding random vectors, like any others.
- As such, these distributions have properties as any other distributions, e.g., means, variances, covariances, etc., which we do not (re-)cover in a systematic fashion but will encounter these properties in the context of our statistical methods.
- We’re very close to **Bayes’ theorem**, but we’ll wait until our context is Bayesian statistics before discussing this theorem.

**Definition 5.6** (Multiplication Rule for Joint pdf/pdf/distribution).

If  $\mathbf{Y} = (Y_1, \dots, Y_n)$  is a random vector with (joint) distribution,  $[\mathbf{y}] = [y_1, \dots, y_n]$ , then

$$[y_1, \dots, y_n] = [y_1][y_2 | y_1][y_3 | y_1, y_2] \cdots [y_n | y_1, y_2, \dots, y_{n-1}],$$

for any ordering of the  $Y_i$ .

- This implies that, if we **have** a joint distribution, then (in principle) it factors into any one of  $n!$  possible products of marginal and conditional distributions. One joint, many possible ways to factor. Gee, that's mildly interesting.
- A more practical implication of the multiplication rule is that we can **build** a joint distribution from a product of marginal and conditional distributions. This seems much more useful and is the essence of **hierarchical modeling** (aka **multi-level modeling**).
- Typically, in practice, we have only one factorization to build a joint distribution, and, generally, it is difficult to create two or more different factorizations that correspond to the same joint!

### Definition 5.7 (Independent Random Variables).

If  $Y_1, \dots, Y_n$  are said to be independent random variables if their joint distribution factors into the product of (scalar) marginal distributions, i.e.,  $\mathbf{Y}$  is a random vector with joint distribution,  $[\mathbf{y}] = [y_1, \dots, y_n]$ , then

$$[y_1, \dots, y_n] = [y_1][y_2][y_3] \cdots [y_n],$$

- This is a special case of the multiplication rule where conditional distributions do not actually depend on the conditioning variables, i.e.,  $[y_i | y_1, \dots, y_{i-1}] = [y_i]$
- Independence is by far the most common way to build a joint distribution as we are often much more comfortable with thinking about distributions of individual random variables rather than of vectors of random variables.
- The joint normal distribution example, above, was obtained via the assumption of independence.

### 5.3 Model Specification

The **traditional linear model** of  $n$  observation pairs,  $(y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , is a model for the conditional distribution,  $[y_i | \mathbf{x}_i]$  arising from

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i,$$

([Wak13, Expr. (5.1)]), i.e.,

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

where

- $Y_i$  is the random variable associated with observation  $y_i$ ,
- $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})^T$  is a vector of observed covariates,
- $\boldsymbol{\beta} = (\beta_0, \dots, \beta_k)^T$  is a vector of parameters,
- and the random **errors**,  $\epsilon_i$  are **uncorrelated** with zero mean,  $E(\epsilon_i) = 0$ , and constant variance across observation  $i$ ,  $\text{Var}(\epsilon_i) = \sigma^2$ .

- As we've seen before, informally, this model can be written in **matrix form** as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

i.e.,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Note, we deviate somewhat from the unconventional notation of [Wak13], who writes (not consistently)  $\mathbf{x}_i$  as a **row vector** and  $\mathbf{x}$  as the **matrix** consisting of **rows** of  $\mathbf{x}_i$ . Our notation is more conventional.
- Figure 5.1 illustrates the linear regression model with regression function  $E(Y | \mathbf{x}) = 10 + 2x_1 + 5x_2$  (Source: [KNNL05]).

### 5.3.1 It's a Conditional Mean Model Specification

- By specifying a distribution only for the **conditional random variable**  $\mathbf{Y} | \mathbf{X}$  (or  $\mathbf{Y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2$  if we want to be explicit about parameters), we are assuming the **joint distribution** factors into the form

$$[\mathbf{y}, \mathbf{x} | \boldsymbol{\beta}, \sigma^2, \gamma] = [\mathbf{y} | \mathbf{x}, \boldsymbol{\beta}, \sigma^2][\mathbf{x} | \gamma]$$

([Wak13, Expr. (5.5)]).

- That is, traditional linear modeling, including "regression and ANOVA," implicitly **assumes that the distribution of the covariates does not depend on, hence does not inform, the**

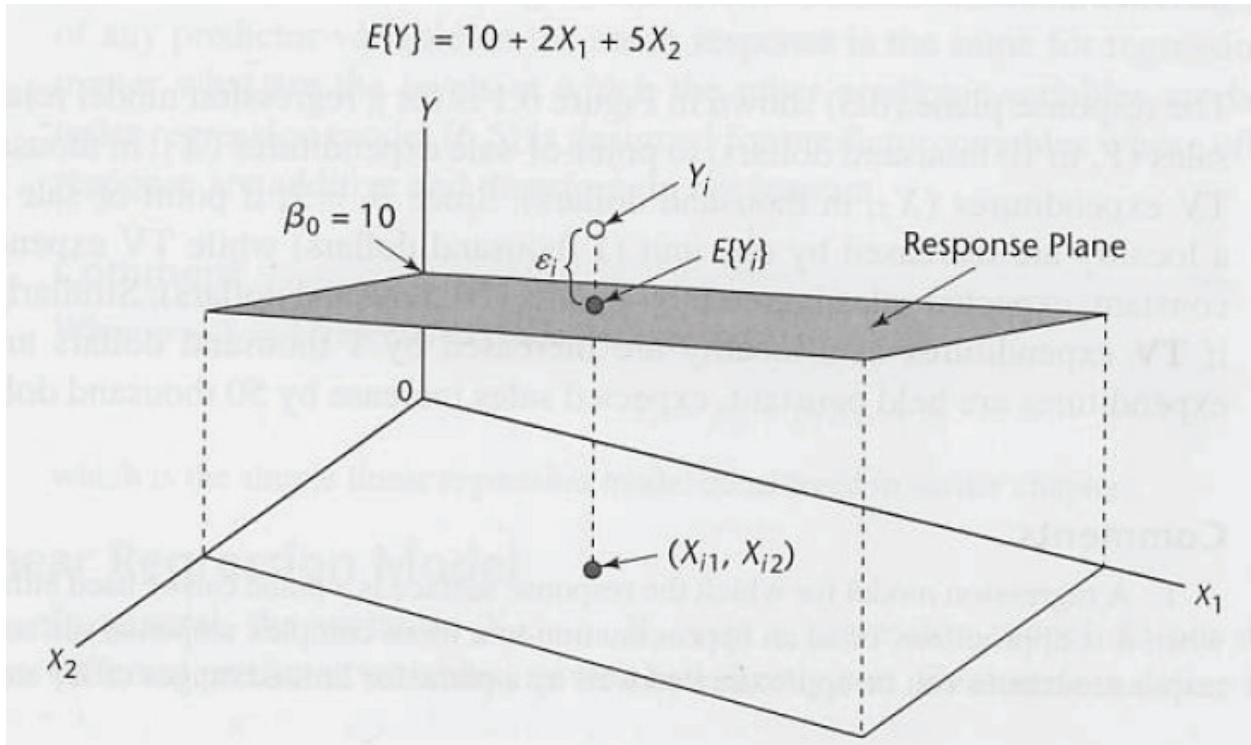


Figure 5.1: Bivariate regression model (Source: [KNNL05]).

**parameters  $\beta$  and  $\sigma^2$** , hence we may ignore the distribution,  $[\mathbf{x} | \gamma]$ , for purposes of inferring  $\beta$  (and  $\sigma^2$ ) (else we may be throwing away information about  $\beta$  and will essentially be estimating a different parameter (despite having the same symbol)).

### Definition 5.8 (Regression Function).

- The **conditional mean** (mean of the conditional distribution or of the conditional random variable) of  $Y | \mathbf{x}$  is denoted as

$$E(Y | \mathbf{x}),$$

and is called the **regression function**.

- Of course, the conditional mean will be of a linear form for much of what we do, at least until we get beyond [Wak13, Chap. 5], i.e.,

$$E(Y | \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

for now.

- Similarly, we may denote the **conditional variance** (variance of the conditional distribution) as

$$\text{Var}(Y | \mathbf{x}) = \sigma^2$$

a relatively uninteresting constant **variance function**. We hope to say more about more interesting variance functions, later.

### 5.3.2 Covariates Observed Without Error

Further, we assume that the process of observing the covariates does not introduce error, i.e., the **covariates are assumed to be observed without error** ([Wak13, p. 198])!

## 5.4 A Justification of Linear Modeling

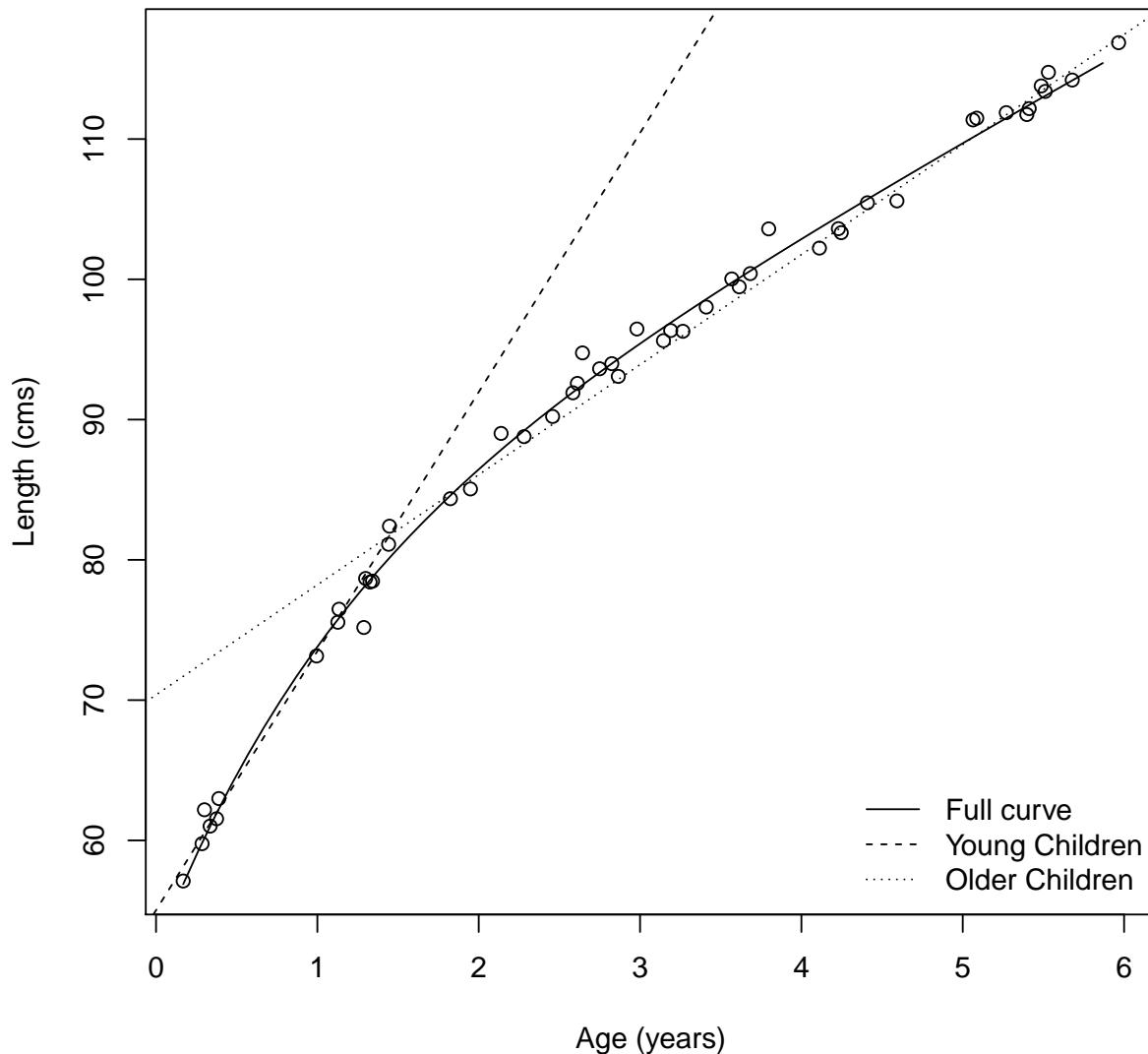
- [Wak13, Chap. 5.4] presents the view that a first order Taylor series approximation of the unknown regression function,  $f(x) = E(Y | x)$  as justification for linear models.

$$\begin{aligned} f(x) &\approx f(x_0) + \left( \frac{df}{dx} \Big|_{x_0} \right) (x - x_0) \\ &\equiv \beta_0 + \beta_1(x - x_0) \end{aligned}$$

- Of course, the approximation may be good only locally, as the next plot of children’s heights (lengths) vs. age suggests ([Wak13, Fig. 5.2]) (the approximating lines in the plot do not quite appear to be tangent lines as the Taylor approximation suggests...).
- **Linear or Linear?** To be sure, “linear” here refers to linearity of the regression function wrt the parameters, not with respect to the covariates. After all, if the regression function truly is linear in the parameters, then we have a linear model as we have been discussing all along—no approximation.
- E.g., Jenss growth curve,

$$\mathbb{E}(Y | x) = \beta_0 + \beta_1 x - \exp(\beta_2 + \beta_3 x),$$

as in the nearby plot.



## 5.5 Parameter Interpretation

### 5.5.1 Conditional Mean Model vs. Marginal Mean model

- As mentioned in the above §5.3, we model the conditional mean, not the marginal mean.

- To illustrate the difference, consider the two models ([Wak13, Exprs. (5.7)& (5.8)]),

$$\begin{aligned} E(Y) &= \beta_0 \quad \text{vs.} \\ E(Y | \mathbf{x}) &= \beta_0. \end{aligned}$$

- The first model says only that the mean of the population of values modeled by the random variable,  $Y$ , is some constant, which does not seem to be a terribly daring or enlightening model!
- In particular, it says nothing about the second model, which, on the other hand, says that the mean of  $Y | \mathbf{x}$  does not vary with  $\mathbf{x}$ , which seems to say quite a lot (though it does not say that the entire distribution of  $Y | \mathbf{x}$  does not vary with  $\mathbf{x}$ ).
- Again, we model the **regression function**, the conditional mean, the second one (with more interesting models than  $\beta_0$ !).

### 5.5.2 Extrapolation, Meaningful Parameters & Reparameterization

- Consider the **line intercept parameter**,  $\beta_0$ , in the **simple linear model (SLR)**
- $$E(Y | x) = \beta_0 + \beta_1 x.$$
- Depending on the data, this parameter **may not make much sense**.
  - For example, if  $y$  is blood serum cholesterol and  $x$  is weight, we may be more interested in the cholesterol of adults whose observed weights are likely clustered away from zero.
  - So, our interest lies far from zero, not near  $x = 0$ , and, further, we do not have data near  $x = 0$  to help inform the relationship between  $y$  and  $x$  in this case.

- Generally, we do not want to infer beyond our data, i.e., we do not want to use our model to **extrapolate**. (Remember, we may view a linear model as a first order approximation of a potentially more complicated function whose behavior is unknown without data.)
- For the children's height vs age data, illustrated in [Wak13, Fig. 5.2], and reproduced in a plot, above, the model makes a bit more sense if we observed data for children whose ages are close to zero. And, presumably, height at age zero is the length of a child at birth, which seems to be a reasonably interesting quantity.

- We may want to **reparameterize** our model to get parameters that are more meaningful.
- For example, a very common reparameterization of the above SLR is

$$E(Y | x) = \beta_0^* + \beta_1(x - x^*),$$

- where  $x^*$  is often chosen as  $x^* = \bar{x}$ , the **average** of the  $x$  observations in the data, or some other value that is a convenient reference level of the covariate.
- For example, **theory** may suggest a value for  $x^*$  where the unknown regression function is known to be approximately linear (as implied by the above discussion about first order Taylor approximation).
- Later, we will see **reference levels of covariates** in the case where a covariate is a categorical variable, as in ANOVA; e.g., perhaps one level of a factor corresponds to a standard treatment or to a placebo, which some find to be a natural reference.

### 5.5.3 Typical “Additive Change” Parameter Interpretation

Here, we give the typical interpretation of linear regression function parameters, but focus on potential mis-interpretations in observational regression studies.

- The common, mathematical interpretation of a linear regression function parameter,  $\beta_k$ , is the **additive change in the expected value of  $Y$  (linear regression function) for a one unit increase in  $x_k$ , all other covariates  $x_j$   $j \neq k$ , held constant.**
- More generally,  $c\beta_1$  is the additive change to the linear regression function with a  $c$  unit (additive) change in  $x$  ([Wak13, p. 200]).
- For example, considering two covariates,  $x_1$  and  $x_2$ , in a multiple linear regression (MLR) model,

$$E(Y | x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

we have the difference

$$\beta_0 + \beta_1(x_1 + c) + \beta_2 x_2 - (\beta_0 + \beta_1 x_1 + \beta_2 x_2) = c\beta_1.$$

- As [Wak13, p.200] warns, we should **be a bit wary of this interpretation** in observational studies.
- It suggests that we could somehow intervene to change every unit's  $x_k$  value in a population by  $c$  to get a corresponding change in the population mean (modeled via our linear regression function),  $c\beta_1$ , which seems to suggest that we can **cause** the mean of  $Y$  to change if we change  $x_k$ , but we know to be careful about such interpretation unless our data come from a randomized experiment. We'll say more about this in the regression context, shortly.
- Also, use of the term, **effect**, to refer to a regression function parameter,  $\beta_k$ , suggests that  $x_k$  and  $Y$  have a **cause and effect** relationship; again, be wary in observational studies, without randomization.

**Example 5.2.** *Confounding in Observational Regression Studies*

- Consider the (unknown) “**true**” relationship,

$$E(Y | x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

([Wak13, Expr. (5.10)], with slight notation change to be consistent with the example on [Wak13, pp. 235-6], to be discussed shortly).

- Assume we are interested in the effect of **observed covariate**,  $x_1$ , on  $Y$ , i.e.,  $\beta_1$ , but we **do not observe**  $x_2$ .
- Also, assume that  $x_1$  and  $x_2$  are **linearly associated** via

$$E(X_2 | x_1) = a + bx_1,$$

([Wak13, Expr. (5.11)], up to slight notation change).

- Now, assume that our **chosen model** is

$$E(Y | x_1) = \beta_0^* + \beta_1^* x_1.$$

- Thus, unobserved  $X_2$  is related to both  $Y$  and to observed covariate,  $X_1$ ; i.e.,  $X_2$  is a **confounder** by our definition (1.17).
- We show that the  $\beta_1^*$  in our **chosen model** is generally not to be considered the same as  $\beta_1$  in the **true model**, without randomization.
- That is, without randomization, we should not think of  $\beta_1^*$  as the effect on the population mean of  $Y$  caused by a unit increase in the covariate  $x_1$  in the population.

- Let’s look at the problem of confounding more clearly in the above example of an observational regression study.

- By the **law of iterated expectations** ([CB02, Theorem 4.4.3])

$$\begin{aligned}
 E(Y | x_1) &= E_{X_2|x_1}(E(Y | x_1, x_2)) \\
 &= E_{X_2|x_1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \\
 &= \beta_0 + \beta_1 x_1 + \beta_2 E(X_2 | x_1) \\
 &= \beta_0 + \beta_1 x_1 + \beta_2(a + bx_1) \\
 &= (\beta_0 + a\beta_2) + (\beta_1 + b\beta_2)x_1 \\
 &= \beta_0^* + \beta_1^* x_1.
 \end{aligned}$$

- The resulting model is the **same as our chosen model**, but now we see more clearly the problem that confounding presents to parameter interpretation without randomization.
- In particular, unless  $X_1$  and  $X_2$  are uncorrelated ( $b = 0$ ) or  $X_2$  does not affect  $Y$  ( $\beta_2 = 0$ ), then  $X_2$  is a confounder, and we see its influence in the **bias**,  $b\beta_2$ , when using  $\beta_1^*$  to infer about the effect of interest,  $\beta_1$ .

### Example 5.3 (Prostate Cancer).

- The example at the end of [Wak13, §5.9] uses the prostate data to illustrate how the bias translates to estimated regression function parameters.
- In particular,  $Y$  is  $\log(\text{PSA})$  (prostate specific antigen) used as a marker to predict cancer;  $x_1$  is  $\log(\text{cap pen})$ , a measure of the extent of the cancer; and  $x_2$  is  $\log(\text{can vol})$ , (cancer volume).

- Estimates are

$$\begin{aligned}
 \hat{\beta}_1 &= 0.66 \quad \text{from fit to true model} \\
 \hat{\beta}_2 &= 0.08 \quad \text{from fit to true model} \\
 \hat{b} &= 0.80 \quad \text{from fit of } a + bx_1 \text{ to } x_2 \\
 \hat{\beta}_1^* &= 0.72 \quad \text{from fit of chosen model} \\
 &= \hat{\beta}_1 + \hat{b} \times \hat{\beta}_2 \\
 &= 0.66 + 0.80 \times 0.08.
 \end{aligned}$$

- Thus, the estimate of the effect of  $\log(\text{cap pen})$  ( $X_1$ ) is inflated (0.72 vs 0.66) due to the estimated effect of  $\log(\text{can vol})$  ( $X_2$ ) on  $\log(\text{cap pen})$  (0.80) and to the estimated effect of  $\log(\text{can vol})$  ( $X_2$ ) on  $Y$  (0.08).

- How would randomization help in these examples of observational regression studies?
- If we could randomly assign units (with their unobserved values of confounder,  $x_2$ , attached) to different levels of  $x_1$ , then there would be **no systematic differences** of units (including their  $x_2$  values) among levels of  $x_1$ .
- In other words,  $E(X_2 | x_1)$  would be the same constant value across  $x_1$  values, i.e.,  $E(X_2 | x_1) = a$  ( $b = 0$ ) so that  $\beta_1^* = \beta_1$ —**no bias**.
- Of course, there would be no opportunity in the prostate example to randomly assign units to values of  $X_1$  ( $\log(\text{cap pen})$ ) values in order to mix-up  $X_2$  ( $\log(\text{can vol})$ ) values!

**Example 5.4** (Lung Cancer, Smoking and Alcohol Consumption).

- *The example on [Wak13, p. 201] illustrates how the association of alcohol consumption ( $x_2$ ) with smoking ( $x_1$ ) can lead to over-estimates of the effect of smoking on cancer ( $Y$ ).*
- *Assuming that alcohol consumption is positively associated with cancer ( $\beta_2 > 0$ ) and with smoking ( $b > 0$ ), then  $\beta_1^*$  reflects effect of smoking and drinking.*
- *In particular, then  $\beta_1^*$  over-“estimates” the effect of smoking on cancer.*
- *So, as the example goes, if we were to intervene—as in an anti-smoking campaign—to decrease smoking by one unit, we do not expect cancer to decrease by  $\beta_1^*$ , but something less (assuming the intervention does not change alcohol consumption, the only confounder).*
- *Of course, akin to the prostate example, it may be difficult to randomly assign people to smoking levels ( $x_1$ ) to mix-up their confounding drinking levels ( $x_2$ )! What about rodents?*

- So, **beware of interpretation of parameters in observational studies.**
- See the recent article in Popular Science about the “benefits” of moderate drinking.
- If we are interested the effect of  $X_1$ , but we also observe  $X_2$  (and perhaps other covariates), which we think may be related to  $Y$  (and perhaps to  $X_1$ ), then these “observed confounders” are sometimes called **control variables** in the sense of controlling for their otherwise confounding effects on parameter interpretation.

- But, including more covariates to control for confounding may have its price, as alluded to in the simulation example of [Wak13, §4.8] illustrating confounding variable adjustment (very similar to our prostate cancer illustration) and **bias-variance trade-off** of a **variable selection** procedure, but it's a bit pre-mature to discuss this now. Perhaps later.

- Sometimes you cannot change one covariate without changing the value of another.
- For example, consider the typical parameterization of a **parabola**:

$$\text{E}(Y | x) = \beta_0 + \beta_1 x + \beta_2 x^2.$$

- In this case, you may want to **reparameterize** to get more meaningful interpretations of parameters (e.g., [Wak13, p. 202]).
- [Wak13, §5.5.2] continues to discuss different parameterizations of linear models, mostly in the context factor variables as in ANOVA, which we treat later.

#### 5.5.4 Data Transformations

- Traditionally, transformations to the response,  $y$ , or to one or more covariates,  $x$ , are performed **to better satisfy assumptions** of the linear model ([Wak13, §5.5.3]).
- For example, the **logarithm** or **square-root** transformations have been popular transformations of  $y$  and  $x$ .
- Typically, transformations to  $y$  are to address **non-constant variance** and/or to satisfy **normality**.

- Transformations to covariates are usually done to get a “better” fitting **mean model** or to get a more meaningful parameter interpretation.
- Because a parameter’s **units** or **scale** depend on those of  $y$  and  $x$ , parameter interpretation will change when variables are transformed.
- A more **modern alternative to transformations** of  $y$  is to maintain  $y$  on its original scale and to focus effort on a **(co-)variance model**, beyond the traditional regression model of constant variance and uncorrelated errors, to capture how the (co-)variance of  $y$  may change with the mean of  $y$ .
- This may be done, indirectly, by introducing **random effects** ([Wak13, Part III], INF512) into the (conditional) mean model and/or adopting a different distributional form for  $y$ , as in **generalized linear models (GLMs)**, to induce a (co-)variance structure, or more directly by explicitly specifying a (co-)variance model ([Wak13, Chaps. 6 & 7], later).
- We may return to data transformations when we consider **model assessment** ([Wak13, §5.10-11]) and **remedial measures** (e.g., [Wak13, §5.6.4]).

# Lecture 6

## Linear Models II: Frequentist Approach

### Contents

---

|       |  |     |
|-------|--|-----|
| 6.1   | General Linear Model . . . . .   | 177 |
| 6.2   | (Ordinary) Least Squares . . . . .   | 179 |
| 6.3   | Gauss–Markov Theorem . . . . .   | 182 |
| 6.3.1 | Remarks on the Gauss Markov Theorem . . . . .                                  | 183 |
| 6.4   | Normal Maximum Likelihood . . . . .  | 184 |
| 6.5   | Fitted Values, Residuals, Hat Matrix and MSE . . . . .                         | 187 |
| 6.6   | Distributions Following from the Normal Linear Model . . . . .                 | 189 |
| 6.6.1 | $z$ and $\chi^2$ Distribution Results . . . . .                                | 189 |
| 6.6.2 | Standard Error of an Estimator . . . . .                                       | 194 |
| 6.6.3 | $t$ and $F$ Distribution Results . . . . .                                     | 195 |
| 6.6.4 | Estimated Standard Error of an Estimator . . . . .                             | 197 |
| 6.6.5 | Summary of Distributional Results . . . . .                                    | 198 |
| 6.7   | Tests and Intervals using $t$ and $F$ Distribution Results . . . . .           | 199 |
| 6.7.1 | Two Basic Questions . . . . .  | 199 |
| 6.7.2 | General Linear Hypothesis . . . . .  | 200 |
| 6.7.3 | $\mathbf{C}\beta$ Approach with $F$ . . . . .                                  | 200 |
| 6.7.4 | Full vs. Reduced Model or Extra Sum-of-Squares Approach . . . . .              | 200 |
| 6.7.5 | Special Case: Omitting Some Variables or Setting Some $\beta_j$ to 0 . . . . . | 202 |
| 6.7.6 | $t$ tests for Scalar $\mathbf{C}\beta$ . . . . .                               | 203 |
| 6.7.7 | $t$ tests for Omitting a Single $x_j$ or Setting $\beta_j = 0$ . . . . .       | 203 |
| 6.7.8 | $t$ -based Confidence Intervals for Scalar $\mathbf{C}\beta$ . . . . .         | 204 |

|  |            |
|--|------------|
| 6.7.9 $t$ -based Prediction Intervals for $Y   \mathbf{x}$ . . . . .                       | 205        |
| <b>6.8 Example Data Analysis Using <math>t</math> and <math>F</math> Results</b> . . . . . | <b>206</b> |
| <b>6.9 Summary and Final Remarks</b> . . . . .   | <b>215</b> |

---

***Main Objectives:***

- General linear model
- Method of ordinary least squares (OLS)
- Gauss-Markov theorem (BLUE)
- Method of (normal) maximum likelihood
- OLS and MLE of linear regression function parameter  $\beta$  is the same
- Typically do not use MLE of  $\sigma^2$
- The assumption of normality in the linear model leads to  $t$  and  $F$  distributions, which underlie much of the classic frequentist inference methods for linear models.
- Tests and intervals using  $t$  and  $F$  distributions
- (Estimated) standard error of the estimator

---

$\mathcal{O}$

**Additional Reading:**

- [Wak13, §5.6]
- [KNNL05, Chap. 1,2,6-8]
- [RS13, Chap. 7, 9 & 10]

---

 $\mathcal{R}$ 

Classic regression and classic ANOVA are special cases of the general linear model and are grouped into the general enterprise of “regression” along with other models of  $E(Y | \mathbf{x})$  (and of  $\text{Var}(Y | \mathbf{x})$ , later, hopefully) by [Wak13]. Here, we cover classic regression from a frequentist perspective. We will get to a ANOVA, later, and to a Bayesian treatment of these linear models.

## 6.1 General Linear Model

Here, we treat the traditional general linear statistical model from a **frequentist perspective**. In particular, for now, we will focus on a few frequentist **goodness criteria** for evaluating estimators of (interesting functions of) the (linear) regression function and for making inference about (interesting functions of) the regression function. In particular, we focus on **maximum likelihood** estimation under the assumption of normal responses and other typical assumptions. Also, we discuss the **least squares** criterion. Later, in other chapters, we *may* look at methods that allow us to relax assumptions about our data while retaining some goodness. Also, we will get to a Bayesian treatment, later, which also depends on the normal likelihood.

- The **general linear model** encompasses previous examples, 4.4, 4.17, 4.20 and 4.22, plus a large part of **machine learning**:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim ?(\mathbf{0}, \boldsymbol{\Sigma}),$$

where

- $\boldsymbol{\Sigma}$  is a general variance (matrix), but,
- for much of what we will do,  $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ , implying the notion that all outputs are of “**equal quality**” (uncorrelated).
- Also, we will more or less begin by assuming that  $\boldsymbol{\epsilon}$  = Normal, so that uncorrelated implies independence.
- Later, hopefully, we will attempt to explore methods that allow “good” inference without a full distributional assumption, usually just a mean and (co-)variance model.
- A **traditional statistical approach** to regression ([Wak13, Except Part IV]) and to linear models in particular ([Wak13, Chap. 5], [KNNL05]), shares much with **machine learning**.
- However, (supervised) machine learning largely focuses on prediction of an “**output**,”  $y$ , at one or more “**inputs**,”  $\mathbf{x}$ , largely to the exclusion of regression function details ([Bis06], [RW06], [HTF09], [Mur12], [Wak13, Part IV], [JWHT14], INF 504).
- In contrast, **in this course**, we focus on estimation of the relationship between  $y$  and one or more  $x$ , usually via interpretable parameters within a (linear) model of the regression function,  $E(Y | \mathbf{x})$ .

## 6.2 (Ordinary) Least Squares

- How do we get a “good” estimator/estimate of the linear model of the regression function, i.e., of the regression function,  $\mu(\mathbf{x})$ ?
- (We merely use  $\mu(\mathbf{x})$  as alternative notation for the regression function, which we have previously denoted as  $E(Y | \mathbf{x})$ .)
- Generally speaking, if we assume that our observed (and unobserved) data are of “equal quality,” then we might think to estimate the regression function,  $\mu(\mathbf{x})$ , using the **method of (ordinary) least squares ((O)LS)**, i.e., choose the particular function, denoted,  $\mu = \hat{\mu}$ , that minimizes the objective function (goodness criterion)

$$RSS(\mu) = \sum_{i=1}^n (Y_i - \mu(\mathbf{x}_i))^2.$$

- But, **infinitely many functions** satisfy this criterion. We need to somehow **regularize** the problem or to provide **more structure** to the problem to find a unique solution.
- In our current context, of course, we impose a linear (in the parameters) model on the regression function, which we may denote as,  $\mu(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}^T \boldsymbol{\beta}$ , to provide structure to (typically) allow us to find a unique value of  $\boldsymbol{\beta}^T = [\beta_0, \dots, \beta_{p-1}]$ , call it  $\hat{\boldsymbol{\beta}}^T = [\hat{\beta}_0, \dots, \hat{\beta}_{p-1}]$ , that minimizes

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^n (Y_i - \mu(\mathbf{x}_i, \boldsymbol{\beta}))^2 \quad [\text{Wak13, §5.6.2}].$$

- (We use  $p$  to denote the number of parameters in our linear regression model. If we have  $k$  covariates, and we include an intercept parameter, then  $p = k + 1$ .)

- This implies “**equal quality**” of responses, which, in our statistical context is embodied by the assumption that  $\text{Var}(Y_i) = \sigma^2$ , constant across  $i$ , i.e.,  $\Sigma = \sigma^2 \mathbf{I}$ .
- Incidentally, we might also write  $RSS$  in **matrix notation**,

$$RSS(\boldsymbol{\beta}) = (\mathbf{Y} - \boldsymbol{\mu})^T (\mathbf{Y} - \boldsymbol{\mu}),$$

- That is, choose the value of  $\boldsymbol{\beta}$  such that the distance (squared) between the observed value,  $\mathbf{y}$ , of the random vector,  $\mathbf{Y}$ , and its mean vector  $\boldsymbol{\mu}$  is minimized.
- Or, more succinctly, **minimize the error/residual sum-of-squares**,

$$RSS(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon}.$$

- That is, choose the value of  $\boldsymbol{\beta}$  to **minimize the squared length of the error vector**. See Figure 6.1 nearby (Source: hover or click here.)
- Upon differentiating with respect to the elements of  $\boldsymbol{\beta}$  and setting equal to zero (why?), we get the **normal equations**

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y} \quad \text{details omitted}$$

(see, e.g., [KNNL05, p. 201], [Wak13, Expr. (5.31)]).

- Almost always in regression—but often not so straight away in ANOVA—we can left multiply by the inverse  $(\mathbf{X}^T \mathbf{X})^{-1}$  to get

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

(see lecture note §4.2.6).

- Statisticians typically refer to  $\hat{\boldsymbol{\beta}}$  as the **estimator** or **estimate** of the unknown parameter  $\boldsymbol{\beta}$ .
- Machine learners might say that  $\hat{\mu} = \mu(\mathbf{x}, \hat{\boldsymbol{\beta}})$  is a “predictor” of output  $y$  given input  $\mathbf{x}$ .

- NOTE: The least squares criterion for finding estimators leads to explicit, **closed-form solutions** for the estimators in the case of linear models. For non-linear (in the parameters!) regression function models,  $\mu(\mathbf{x})$ , the least squares criterion typically only implicitly defines estimators and values of estimator must be found by some iterative algorithm.

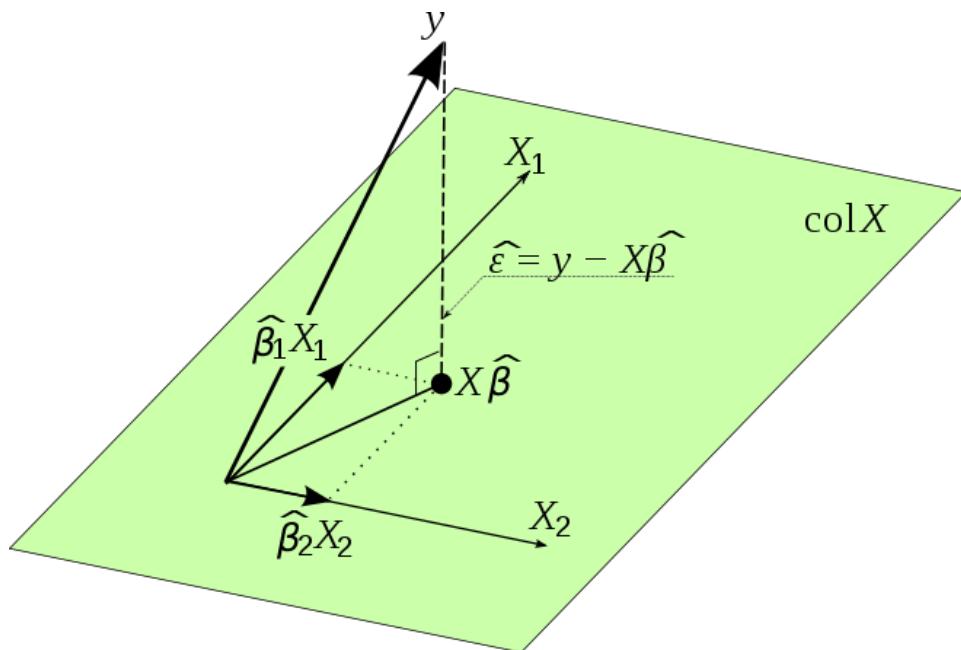


Figure 6.1: Geometric interpretation of OLS for linear models. (Source: see text.)

### 6.3 Gauss–Markov Theorem

- Somehow, minimizing (the sum-of-squared) errors feels good. Naturally, we want more goodness! What are other ways in which we might view the “goodness” of the (O)LS estimator  $\hat{\beta}$ ? Are there other evaluation criteria?
- According to our linear model (without the normal assumption yet), let

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \quad \epsilon \sim ?(\mathbf{0}, \sigma^2 \mathbf{I}),$$

and consider as above the estimator

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

- Then,  $\hat{\beta}$  is called the **Best Linear Unbiased Estimator of  $\beta$  (BLUE)**.
- Again, we don’t need to assume  $\epsilon \sim N$  for this result, but we do assume mean zero and variance  $\sigma^2 \mathbf{I}$ , though a very similar version of BLUEness that holds for a general variance matrix  $\Sigma$ , with a caveat.

- **Best** in the following sense (slightly more general than in [Wak13, §5.6.3]). Let

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix},$$

and let

$$\mathbf{c} = [c_0, \dots, c_{p-1}]$$

be any single row matrix of  $p$  constants. Then,

$$\text{Var}(\mathbf{c}\hat{\beta}) \leq \text{Var}(\mathbf{c}\hat{\beta}^*),$$

where  $\widehat{\boldsymbol{\beta}}^*$  is any estimator you can create of the form  $\widehat{\boldsymbol{\beta}}^* = \mathbf{a} + \mathbf{A}\mathbf{Y}$  with  $E(\widehat{\boldsymbol{\beta}}^*) = \boldsymbol{\beta}$ . That is, if you consider any linear unbiased estimator of  $\boldsymbol{\beta}$ , then linear combinations of  $\widehat{\boldsymbol{\beta}}$  have at least as small variances as the same linear combinations of your estimator.

- **Linear** means the estimator is a linear function of the data, i.e., is of the form  $\mathbf{a} + \mathbf{B}\mathbf{Y}$  (see §4.3.3 & 4.3.4).
- **Unbiased** means, as we have seen,  $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ .
- **Estimator** means it is a function of our (random) data vector  $\mathbf{Y}$  and, in particular, does not depend on  $\boldsymbol{\beta}$ . (It would be unfortunate if it depended on the unknown quantity being estimated!)

- We will see plenty of realistic examples of  $\mathbf{c}$  and of matrices  $\mathbf{C}$  consisting of rows of such vectors.

### 6.3.1 Remarks on the Gauss Markov Theorem

- A more general version of GM assumes

$$\boldsymbol{\epsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma}),$$

for general variance matrix  $\boldsymbol{\Sigma}$ . In this case, the “good” estimator is  $\widehat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y}$ , called the **generalized least squares (GLS)** estimator of  $\boldsymbol{\beta}$  ([Wak13, §5.6.2]).

- In this more general case,  $\boldsymbol{\Sigma}$  is assumed known, but in practice we almost never know it! It’s an unknown parameter (matrix)! If we instead plug in an estimate for it, then this generalized GM theorem is technically **out the window**.

- But, as we've seen, special forms of  $\Sigma$  allows it to be unknown, and we still get bestness from GM. This includes  $\Sigma = \sigma^2 \mathbf{I}$ , which is the form we will use for much of what we do. We still have "bestness"!

- If, in addition to the conditions of the GM, we assume further that  $\epsilon$  is normal, then  $\hat{\beta}$  is not only BLUE, it is BUE (Best Unbiased Estimator: best among all functions of the data, linear (in the parameters) or not).
- However, BLUE may be far worse than BUE if  $\epsilon$  is not normal.
- The goodness of GM depends on our linear statistical model being the **true model** (again, normality is required for BUE but not for BLUE.)
- In **machine learning**, we often focus on a search for the **best predictive model**, whether it be linear, non-linear, explicit or implicit. While LS (and ML and other) methods remain relevant, ultimately, there is typically one goodness criterion, more directly related to prediction (rather than to estimation of  $\beta$  as in LS and ML), that guides the search for a model.
- For the more predictive perspective, take INF 504 ([Bis06], [RW06], [HTF09], [Mur12], [Wak13, Part IV], [JWHT14]).

## 6.4 Normal Maximum Likelihood

**Maximum likelihood** (ML) is another goodness criterion used to estimate  $\mu$  (and  $\sigma^2$ ). We will essentially begin with the normal (or other, e.g., Binomial or Poisson) likelihood then seek to relax some of the assumptions underlying the use of likelihoods, later.

- Assuming normality,  $\epsilon \sim N(0, \sigma^2)$ , then  $y \sim N(\mu(\mathbf{x}, \boldsymbol{\beta}), \sigma^2)$ , and the (normal) **likelihood function** of  $n$  **independent** observations is ([Wak13, §5.6.1])

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2}\sum_i \left(\frac{y_i - \mu(\mathbf{x}, \boldsymbol{\beta})}{\sigma}\right)^2\right),$$

and the (estimation) **method of maximum likelihood** tells us to choose the values of the parameters,  $\boldsymbol{\beta}$  and  $\sigma^2$ , that maximize the likelihood function,  $L(\boldsymbol{\beta}, \sigma^2)$  (with the data fixed at their observed values). That is, what value of the parameters make the data “most likely” inasmuch as the likelihood function represents “likeliness”?

- See Example 3.1 and Definition 3.3 (for the pdf of a single normal random variable) and Definitions 5.6 and 5.7 (multiplication rule and independence, respectively).
- In **matrix-vector notation**,

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{x}^T \boldsymbol{\beta})^T(\mathbf{y} - \mathbf{x}^T \boldsymbol{\beta})\right).$$

- (We write  $\mathbf{y}$  instead of  $\mathbf{Y}$  because the likelihood function is often considered conditional on observed data,  $\mathbf{y}$ , though sometimes (functions of) the likelihood consider the random version,  $\mathbf{Y}$  (e.g., score function) to derive properties of maximum likelihood estimators.)
- Equivalently, we can find the value of the parameters that maximize the (normal) **log-likelihood function** ([Wak13, §5.6.1])

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2} \sum_i \left(\frac{y_i - \mu(\mathbf{x}, \boldsymbol{\beta})}{\sigma}\right)^2,$$

and we see that maximizing the (normal) likelihood in  $\boldsymbol{\beta}$ , for any given  $\sigma > 0$ , is equivalent to the aforementioned LS criterion of minimizing the sums of squared errors.

- In **matrix-vector notation**,

$$L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2}(\mathbf{y} - \mathbf{x}^T \boldsymbol{\beta})^T (\mathbf{y} - \mathbf{x}^T \boldsymbol{\beta}).$$

- **Maximum likelihood estimators/estimates (MLEs)** follow by finding the values of the parameters ( $\boldsymbol{\beta}$  and  $\sigma$ ) that maximize the (log-)likelihood function (after a bit of calculus, in our current case ([Wak13, p. 209]))

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \\ \hat{\sigma}^2 &= \frac{1}{n} (\mathbf{Y} - \mathbf{x}^T \hat{\boldsymbol{\beta}})^T (\mathbf{Y} - \mathbf{x}^T \hat{\boldsymbol{\beta}}) \\ &= \frac{1}{n} \hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}\end{aligned}$$

where we have defined

$$\hat{\boldsymbol{\epsilon}} = (\mathbf{Y} - \mathbf{x}^T \hat{\boldsymbol{\beta}});$$

see Definition 6.2 of **residuals**, below.

- The above MLE of  $\sigma^2$  is biased, and we often use an unbiased estimator of  $\sigma^2$ ,

$$\hat{\sigma}^2 = \frac{1}{n-p} = (\mathbf{Y} - \mathbf{x}^T \hat{\boldsymbol{\beta}})^T (\mathbf{Y} - \mathbf{x}^T \hat{\boldsymbol{\beta}});$$

see Definition 6.3 of **MSE**, below.

- Note, we (and [Wak13]) often do not use notation to distinguish the MLE from the unbiased estimator.
- Our **square-bracket notation** (Definition 5.3) in the context of the normal linear model for the joint distribution of observed random variables,  $Y_i$ , may be written

$$[\mathbf{y} | \boldsymbol{\beta}, \boldsymbol{\Sigma}] = N(\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta}), \boldsymbol{\Sigma}),$$

or, for a single observation,

$$[y | \boldsymbol{\beta}, \sigma^2] = N(\mu(\mathbf{x}, \boldsymbol{\beta}), \sigma^2),$$

## 6.5 Fitted Values, Residuals, Hat Matrix and MSE

Now is a fair time to introduce more notation and definitions in terms of the linear model. Refer to [Wak13, §5.11.2] and [KNNL05, Sec. 6.4]. [RS13] tend to minimize notation, but see [RS13, Sec. 7.3] in the SLR context. Similar concepts are used for other regression models to, which we hope to see later.

**Definition 6.1** (Fitted Values).

$$\begin{aligned}\hat{\mathbf{Y}} &= (\hat{Y}_1, \dots, \hat{Y}_n)^T \\ &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\text{"H" at matrix}}\mathbf{Y} \\ &= \mathbf{HY}\end{aligned}$$

or, *observation-wise*,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_{p-1} x_{ip-1}.$$

**Definition 6.2** (Residuals).

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}} &= (\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^T \\
&= \mathbf{Y} - \hat{\mathbf{Y}} \\
&= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{Y} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \\
&= (\mathbf{I} - \underbrace{\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T}_{\text{"H" at matrix}}) \mathbf{Y} \\
&= (\mathbf{I} - \mathbf{H}) \mathbf{Y}
\end{aligned}$$

([Wak13, Expr. (5.60)], where his  $\mathbf{h}$  is our  $\mathbf{H}$ ) or, observation-wise,

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i,$$

([Wak13, Expr. (5.59)]). Later we will use the hat matrix  $\mathbf{H}$  and residuals for diagnosing model assumptions ([Wak13, §5.11.2], [RS13, Sec. 11.4] and [KNNL05, Chap. 10]).

$$\begin{aligned}
Var(\hat{\boldsymbol{\epsilon}}) &= Var((\mathbf{I} - \mathbf{H}) \mathbf{Y}) \\
&= (\mathbf{I} - \mathbf{H}) \sigma^2 \mathbf{I} (\mathbf{I} - \mathbf{H})^T \\
&= \sigma^2 (\mathbf{I} - \mathbf{H})^2 \quad (\text{symmetric}) \\
&= \sigma^2 (\mathbf{I} - \mathbf{H}) \quad (\text{idempotent})
\end{aligned}$$

Observation-wise, we have

$$Var(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii}),$$

where  $h_{ii}$  is the  $i$ th diagonal element of  $\mathbf{H}$ . Again, see [Wak13, §5.11.1], and he uses notation  $\mathbf{h}$  in place of our  $\mathbf{H}$ .

**Definition 6.3** (Unbiased Estimator of  $\sigma^2$  in a Linear Model (MSE)).

$$\begin{aligned}
MSE &= \hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n-p} = \frac{\sum_{i=1}^n (\hat{\epsilon}_i - \bar{e})^2}{n-p} \\
&= \hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}} / (n-p)
\end{aligned}$$

where, now,  $p \geq 2$ , and  $\bar{e} = \mathbf{0}$  is the average residual value (it's **always zero**; used here explicitly only to convey that MSE looks very much like a typical empirical variance, with slightly different denominator).

## 6.6 Distributions Following from the Normal Linear Model

### 6.6.1 $z$ and $\chi^2$ Distribution Results

- Perhaps feeling good about GM, or about minimizing error or about maximizing likelihood, statisticians proceed to conduct statistical inference.
- In particular, **for now, we will assume a normal likelihood**, which we hope to relax, later (as I seem to be repeating, a lot.)
- These results presented here are more or less contained in [Wak13, pp. 212-3], which also contains other results under less stringent assumptions, which, as mentioned, we hope cover, later (!). We give many more details here than [Wak13, pp. 212-3].
- Great. But, how do we conduct inference? Patience.
- As in our lecture chapter 1, we seek an estimator and its distribution. As mentioned there, we **assume of normality** (for now) for  $\mathbf{Y}$ , perhaps supported by the **CLT** (if  $n$  is reasonably large) or a **histogram** (of residuals), etc.

- As indicated above,  $\hat{\beta}$  is a linear function of the data (random) vector,  

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

- By previous results for a linear function of a (normal) random vector (§4.3.3 4.3.4 and 4.3.5) we obtain
- the **mean** (vector) of our estimator,

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\mathbf{Y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta}$$

(unbiased for  $\boldsymbol{\beta}$  by the assumption that our linear regression model is correct),

- and its **variance** (matrix)

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \text{Var}(\mathbf{Y}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbf{I} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}, \end{aligned}$$

### **Result 6.1** (Normal Distribution for $\hat{\boldsymbol{\beta}}$ ).

*Combined with the estimator's mean and variance, above, the result of §4.3.5 gives the **normal distribution** for our estimator*

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}).$$

- We state a more generally useful result before explaining how these results are used to make inference.

### **Result 6.2** (Normal Distribution for Linear Combination $\mathbf{C}\hat{\boldsymbol{\beta}}$ ).

By the same results used above (§4.3.3, 4.3.4 and 4.3.5), we have

$$\mathbf{C}\hat{\boldsymbol{\beta}} \sim N(\mathbf{C}\boldsymbol{\beta}, \sigma^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T),$$

for  $\mathbf{C}$  a matrix of one or more known row vectors (linear combination coefficients).

### Result 6.3 (Standard Normal Distribution for Standardized $\mathbf{C}\hat{\boldsymbol{\beta}}$ ).

By the same results used above (§4.3.3, 4.3.4 and 4.3.5), we have

$$\sqrt{\sigma^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta}) \sim N(\mathbf{0}, \mathbf{I}).$$

We use the notation  $\sqrt{\sigma^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T}$  do denote a “**matrix square-root**”: For any symmetric, non-negative definite matrix  $\mathbf{A}$ , there is a (square, symmetric, nnd) matrix  $\mathbf{R}$  such that  $\mathbf{A} = \mathbf{R}^2$  ([Har97, Theorem 21.9.1]); there are other types of matrix square roots (more or less) that may be more popular ([Har97, §14.5]). Incidentally, in practice, most variance matrices have such a factorization into the product of “square-root” matrices.

### Result 6.4 ( $\chi^2(df=\text{rank}(\mathbf{C}))$ for Quadratic Form of $\mathbf{C}\hat{\boldsymbol{\beta}}$ ).

- If we take the standard normal vector of Result 6.3 and take the inner product with itself, so that we get a sum-of-squared standard normals, then we get a  $\chi^2(df=\text{rank}(\mathbf{C}))$ .
- That is,

$$(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta})^T (\text{Var}(\mathbf{C}\hat{\boldsymbol{\beta}}))^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta}) \sim \chi^2(\text{rank}(\mathbf{C})),$$

or, more precisely,

$$(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta})^T (\sigma^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta}) \sim \chi^2(\text{rank}(\mathbf{C})).$$

- We talked about the **rank** of a matrix in Definition 4.7 and §4.2.6. Typically, in the context of linear model practice, it refers to the number of rows of  $\mathbf{C}$ , i.e., the number of linear combinations of  $\boldsymbol{\beta}$  about which we wish to (simultaneously) infer.

**Result 6.5** ( $N(0,1)$  for Standardized Scalar  $\mathbf{C}\hat{\boldsymbol{\beta}}$ ).

- If  $\mathbf{C}\boldsymbol{\beta}$  is a scalar (i.e., if  $\mathbf{C}$  is a single row), then Result 6.3 gives a single  $N(0, 1)$  rv.
- That is,

$$\frac{\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta}}{\sqrt{\text{Var}(\mathbf{C}\hat{\boldsymbol{\beta}})}} \sim N(0, 1),$$

or, more precisely,

$$\frac{\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta}}{\sqrt{\sigma^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T}} \sim N(0, 1).$$

- We may perform hypothesis tests and construct confidence intervals using this reference distribution (if we know  $\sigma^2$ ).

- For example, we may want to infer about the linear regression function at a particular covariate value, i.e., about  $\mathbf{x}^T \boldsymbol{\beta}$  (in this case  $\mathbf{C}$  is just the row vector  $\mathbf{x}^T$ ).

**Result 6.6** ( $N(0,1)$  for Standardized  $\hat{\beta}_j$ ).

- An important special case of Result 6.5 is when  $\mathbf{C}$  is a row vector consisting of a single element of 1 and remaining elements 0.
- That is,

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2(\mathbf{X}^T \mathbf{X})_{(jj)}^{-1}}} \sim N(0, 1),$$

where  $(\mathbf{X}^T \mathbf{X})_{(jj)}^{-1}$  denotes the diagonal element of  $(\mathbf{X}^T \mathbf{X})^{-1}$  corresponding to parameter  $\beta_j$ .

- $(Var(\widehat{\beta})) = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$  is just the variance matrix of the  $\widehat{\beta}$  vector, and we are merely picking off the correct variance element from its diagonal.)
- We may perform hypothesis tests and construct confidence intervals using this reference distribution (if we know  $\sigma^2$ ).

### **Result 6.7** ( $\chi^2(df = 1)$ for Inferring $\widehat{\beta}_j$ ).

- Squaring the above standard normal in Result 6.6 we get a  $\chi^2(df=1)$ .
- That is,

$$\frac{(\widehat{\beta}_j - \beta_j)^2}{Var(\widehat{\beta}_j)} \sim \chi^2(df = 1),$$

or, more precisely,

$$\frac{(\widehat{\beta}_j - \beta_j)^2}{\sigma^2(\mathbf{X}^T \mathbf{X})_{(jj)}^{-1}} \sim \chi^2(df = 1).$$

- We will not use this result or the more general  $\chi^2$  Result 6.4 much, but may refer to it when discussing the (asymptotic) distribution of the likelihood ratio statistic ([Wak13, §2.9.5]) and the relationship of

*a  $\chi^2$  random variable to an  $F$  random variable, which we will use a lot (just below).*

- Of course, in the above series of results, we do not know the parameters  $\beta$  or  $\sigma^2$ .
- **How would we use these distributional results as reference distributions for statistical inference?**
- The unknown  $\beta$  is not a problem, we will typically **hypothesize a value** for it (in testing) or would use the results to obtain an interval estimate of  $\beta$  (or  $C\beta$ ) whose **endpoints do not depend on the unknown  $\beta$** .
- Still, our inferences (tests or intervals) **depend on knowing  $\sigma^2$** , which is not realistic in many cases.

### 6.6.2 Standard Error of an Estimator

**Result 6.8** (Standard Error of an Estimator).

- If  $\hat{\theta}$  is a scalar estimator, then the standard error

$$SE(\hat{\theta}) \equiv \sqrt{Var(\hat{\theta})}.$$

- In other words, the **standard error** of an estimator is just another name for the **standard deviation** of an estimator.

- Typically, the standard error (standard deviation) is discussed only for scalars, as the square root of their variance, but you could define a standard error (standard deviation) square root matrix. (Not common.)
- For us, in the scalar case of  $\mathbf{C}\hat{\boldsymbol{\beta}}$  (Result 6.5),

$$\text{SE}(\mathbf{C}\hat{\boldsymbol{\beta}}) = \sqrt{\sigma^2 \mathbf{C}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}}.$$

- And, more particularly, in the case of  $\hat{\beta}_j$  (Result 6.6),

$$\text{SE}(\hat{\beta}_j) = \sqrt{\sigma^2 (\mathbf{X}^T \mathbf{X})_{(jj)}^{-1}}.$$

### 6.6.3 $t$ and $F$ Distribution Results

- Upon replacing  $\sigma^2$  with  $\hat{\sigma}^2 = \text{MSE}$  (Definition 6.3) in the above results on normal and  $\chi^2$ -square random vectors/variables, we get analogous results for  $t$  and  $F$  random variables, which do not depend on  $\sigma^2$ , and which are more practically useful for inference.
- Again, as mentioned above, the MLE and LS estimators of  $\boldsymbol{\beta}$  are the same in the case of the traditional normal linear model, but the unbiased MSE estimator of  $\sigma^2$  is not the same as the MLE estimator.

**Result 6.9** ( $t(df = n - p)$  for Standardized Scalar  $\mathbf{C}\hat{\boldsymbol{\beta}}$ ).

- If  $\mathbf{C}\boldsymbol{\beta}$  is a scalar (i.e., if  $\mathbf{C}$  is a single row), then, upon replacing  $\sigma^2$  (in  $\text{Var}(\mathbf{C}\hat{\boldsymbol{\beta}})$ ) in the normal Result 6.5 with our unbiased estimate  $\hat{\sigma}^2$ , we get a  $t$  random variable.

- That is,

$$\frac{\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta}}{\sqrt{\widehat{\text{Var}}(\mathbf{C}\hat{\boldsymbol{\beta}})}} \sim t(df = n - p),$$

or, more precisely,

$$\frac{\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T}} \sim t(df = n - p).$$

**Result 6.10** ( $t(df = n - p)$  for Standardized  $\hat{\beta}_j$ ).

- Similarly, and more particularly, upon replacing  $\sigma^2$  in the normal Result 6.6 with our unbiased estimate  $\hat{\sigma}^2$ , we get a  $t$  random variable.
- That is,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})_{(jj)}^{-1}}} \sim t(df = n - p),$$

where, again,  $(\mathbf{X}^T \mathbf{X})_{(jj)}^{-1}$  denotes the diagonal element of  $(\mathbf{X}^T \mathbf{X})^{-1}$  corresponding to parameter  $\beta_j$ .

- $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}$  is just the estimated variance matrix of the  $\hat{\boldsymbol{\beta}}$  vector, and we are merely picking off the correct variance element from its diagonal.)

**Result 6.11** ( $F(df_1 = \text{rank}(\mathbf{C}), df_2 = n - p)$  for Quadratic Form of  $\mathbf{C}\hat{\boldsymbol{\beta}}$ ).

- Replacing  $\sigma^2$  (in  $\text{Var}(\mathbf{C}\hat{\boldsymbol{\beta}})$ ) in the  $\chi^2$  Result 6.4 with our unbiased estimate  $\hat{\sigma}^2$ , we get a  $F(df_1 = \text{rank}(\mathbf{C}), df_2 = n - p)$  random variable.

- That is,

$$(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta})^T (\widehat{\text{Var}}(\mathbf{C}\hat{\boldsymbol{\beta}}))^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta}) \sim F(df_1 = \text{rank}(\mathbf{C}), df_2 = n-p),$$

or, more precisely,

$$(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta})^T (\hat{\sigma}^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{C}\boldsymbol{\beta}) \sim F(df_1 = \text{rank}(\mathbf{C}), df_2 = n-p),$$

- Note that  $t^2(n-p) = F(1, n-p)$ , but we typically use  $t$  results when  $\text{rank}(\mathbf{C}) = 1$ , as in Results 6.9 and 6.10.
- These  $t$  and  $F$  results are used repeatedly throughout linear models (linear regression and ANOVA).
- Again, they depend on our linear model assumptions, including ASUMPTION OF NORMAL ERRORS.

#### 6.6.4 Estimated Standard Error of an Estimator

**Result 6.12** (Estimated Standard Error of an Estimator).

- If  $\hat{\boldsymbol{\theta}}$  is a scalar estimator, then the estimated standard error

$$\widehat{SE}(\hat{\boldsymbol{\theta}}) \equiv \sqrt{\widehat{\text{Var}}(\hat{\boldsymbol{\theta}})}.$$

- In other words, the **estimated standard error** of an estimator is just another name for the **estimated standard deviation** of an estimator.

- Typically, the standard error (standard deviation) is discussed only for scalars, as the square root of their variance, but you could define a standard error (standard deviation) square root matrix. (Not common.)
- For us, in the scalar case of  $\mathbf{C}\hat{\beta}$  (Result 6.9),

$$\widehat{\text{SE}}(\mathbf{C}\hat{\beta}) = \sqrt{\hat{\sigma}^2 \mathbf{C}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}}.$$

- And, more particularly, in the case of  $\hat{\beta}_j$  (Result 6.10),

$$\widehat{\text{SE}}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})_{(jj)}^{-1}}.$$

### 6.6.5 Summary of Distributional Results

Thus, given the assumptions of the ***normal*** linear model, we have several distributions, related to  $\hat{\beta}$ , to help us conduct inference, i.e., to compute **p-values** and obtain **confidence intervals**. Usually, under the assumption of normality, we work with  $t$  and  $F$ , though  $z$  and  $\chi^2$  may sometime be used, especially when relaxing the normality assumption and relying on asymptotic results (later hopefully).

Much of the material presented here will serve us later when we cover Bayesian linear models. Not only will the same quantities appear (i.e., same matrices and vectors), but also the same inferences, in some sense, in some cases.

Also, to be sure, the results here apply to the normal linear model, including linear regression and ANOVA. We tend to illustrate with regression first, then ANOVA, later.

## 6.7 Tests and Intervals using $t$ and $F$ Distribution Results

### 6.7.1 Two Basic Questions

1. Is the parameter this? e.g., is  $\theta = \theta_0$  or is it some alternative value?
2. What is the parameter  $\theta$ ?

We use a hypothesis (or significance) test to help us answer the first type of question, and we use an (interval) estimate to help us answer the second type of question, using statistics from our sample(s) in both cases, of course.

Typically, we obtain interval estimates for scalar quantities, like  $\beta_j$ , using a  $t$  distribution (Result 6.10), but, in principle, we could obtain confidence regions for vectors, e.g., for  $\mathbf{C}\beta$  via an  $F$  distributions (Result 6.11), though this is not commonly done.

The mechanistic, plug-n-chug essence of a (two-sided) **confidence interval** is

$$\text{estimate} \pm \text{multiplier} * (\text{estimate of variability of estimator})$$

and, for the **testing** situation, we compare

$$\text{test-statistic} = \frac{(\text{estimate} - \text{hypothesized parameter value})}{(\text{estimate of variability of estimator})}$$

to some reference distribution to see if our computed test statistic is extreme relative to some null distribution—we'll get to this shortly.

### 6.7.2 General Linear Hypothesis

Consider the **General Linear Hypothesis**

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b}_0 \quad \text{null hypothesis}$$

$$H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{b}_0 \quad \text{alternative hypothesis}$$

for some matrix of linear combination coefficients  $\mathbf{C}$ . We will not consider hypotheses/estimation involving non-linear functions of  $\boldsymbol{\beta}$ , at least not using frequentist methods; otherwise we might appeal to the delta method ([Wak13, Appendix G]).

### 6.7.3 $\mathbf{C}\boldsymbol{\beta}$ Approach with $F$

In this case, we may test the hypothesis via an  $F$  statistic (Result 6.11) under the null hypothesis,

$$F = (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{b}_0)^T (\hat{\sigma}^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{b}_0) \quad (6.1)$$

$$\sim F(df_1 = \text{rank}(\mathbf{C}), df_2 = n - p). \quad (6.2)$$

Given our data (in the estimates  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$ ), along with the null value,  $\mathbf{b}_0$ , we can compute  $F$ , then compare it to the distribution of  $F(df_1 = \text{rank}(\mathbf{C}), df_2 = n - p)$ ,  $p = k + 1$ , i.e., compute a p-value.

### 6.7.4 Full vs. Reduced Model or Extra Sum-of-Squares Approach

An equivalent way to implement this test is to recognize the **null hypothesis** to correspond to a restricted or **reduced model**, and the **alternative hypothesis** to correspond to an unrestricted or **full model**. From this perspective, let

$$\begin{aligned}
 RSS(\boldsymbol{\beta}_F) &= (\mathbf{Y} - \mathbf{x}^T \hat{\boldsymbol{\beta}}_F)^T (\mathbf{Y} - \mathbf{x}^T \hat{\boldsymbol{\beta}}_F) \\
 &= \hat{\boldsymbol{\epsilon}}_F^T \hat{\boldsymbol{\epsilon}}_F \\
 &= (n - p) \hat{\sigma}_F^2
 \end{aligned}$$

and

$$\begin{aligned}
 RSS(\boldsymbol{\beta}_R) &= (\mathbf{Y} - \mathbf{x}^T \hat{\boldsymbol{\beta}}_R)^T (\mathbf{Y} - \mathbf{x}^T \hat{\boldsymbol{\beta}}_R) \\
 &= \hat{\boldsymbol{\epsilon}}_R^T \hat{\boldsymbol{\epsilon}}_R \\
 &= (n - p_R) \hat{\sigma}_R^2.
 \end{aligned}$$

We use the subscripts  $F$  and  $R$  to refer to the **full model**—corresponding to the unrestricted parameters of the **alternative hypothesis**—and to the **reduced model**—corresponding to the restricted parameters under the **null model**—respectively. In general, it may not be easy to see the correspondence between the full model and the alternative or between the reduced model and the null, but often the correspondence is easy to see, as will illustrate soon enough.

If the full model **residual sum-of-squares**,  $RSS(\boldsymbol{\beta}_F)$ , is much less than reduced model's,  $RSS(\boldsymbol{\beta}_R)$ , i.e., if

$$RSS(\boldsymbol{\beta}_R) - RSS(\boldsymbol{\beta}_F)$$

is somehow “large,” then we may tend to think that the reduced model does not “explain” (beware cause-effect interpretation) as much of the residual variability as does the full model, and we may tend to reject the reduced model in favor of the full model. This intuition is good, but we standardize by the difference in the number of parameters and compare the difference relative to that of the full model. That is, we will consider the “largeness” of

$$\frac{(RSS(\boldsymbol{\beta}_R) - RSS(\boldsymbol{\beta}_F))/(p_F - p_R)}{RSS(\boldsymbol{\beta}_F)/(n - p_F)}, \quad (6.3)$$

where  $p_F = p$  and  $p_R$  is the number of free linear regression function parameters in the reduced model (skipping details here). Without showing details,

it turns out that

$$\begin{aligned} F &= (\mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}_0)^T (\hat{\sigma}_F^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}}_F - \mathbf{b}_0) \\ &= \frac{(RSS(\boldsymbol{\beta}_R) - RSS(\boldsymbol{\beta}_F))/(p_F - p_R)}{RSS(\boldsymbol{\beta}_F)/(n - p_F)}, \\ &= \frac{(RSS(\boldsymbol{\beta}_R) - RSS(\boldsymbol{\beta}_F))/(rank(\mathbf{C}))}{RSS(\boldsymbol{\beta}_F)/(n - p_F)} \end{aligned}$$

( $rank(\mathbf{C}) = p_F - p_R$ ). We will illustrate with a few examples as we go. To be sure, in our  $F$  statistic in Equation (6.3), above, we were implicitly in the context of the full model only, so we didn't need distinguishing subscripts.

Thus, we can

1. compute  $F$  via Equation (6.1), or
2. fit a full model and a reduced model and compute  $F$  with Equation (6.3), where, again,  $p_F - p_R = rank(\mathbf{C})$ .

In practice, it is often easy to determine  $rank(\mathbf{C})$  as simply the number of (linearly independent) rows of  $\mathbf{C}$ , or  $p_R$  is often obvious so that  $p_F - p_R$  is easy.

### 6.7.5 Special Case: Omitting Some Variables or Setting Some $\beta_j$ to 0

Frequently, we have a particular form of hypothesis in mind where  $\mathbf{C}\boldsymbol{\beta}$  corresponds to setting some subset of  $q$  parameters to zero ([Wak13, p. 212]). In this case,

- What is  $\mathbf{C}$  (say  $k = 4$  covariates, and we want to know if we can omit  $x_2$  and  $x_4$ )?
- What is  $\mathbf{b}_0$ ?
- What is  $rank(\mathbf{C})$ ?
- What is  $p_F$ ?  $p_R$ ?

The so-called **overall  $F$ -test** is of this form; we leave this to a subsequent chapter of notes.

### 6.7.6 $t$ tests for Scalar $\mathbf{C}\beta$

In particular, for scalar  $t$  Result 6.9 we have

$$H_0 : \mathbf{C}\beta = b_0 \quad \text{null hypothesis}$$

$$H_1 : \mathbf{C}\beta \neq b_0 \quad \text{alternative hypothesis (or } > \text{ or } <\text{)}$$

where  $\mathbf{C}$  is now a **row matrix** and the null hypothesized value,  $b_0$ , is a scalar.

And, we may test the hypothesis via a  $t$  statistic under the null hypothesis,

$$\frac{\mathbf{C}\hat{\beta} - b_0}{\sqrt{\hat{\sigma}^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T}} \sim t(df = n - p).$$

Given our data (in the estimates  $\hat{\beta}$  and  $\hat{\sigma}^2$ ), along with the null value,  $b_0$ , we can compute  $t$ , then compare it to the distribution of  $t(df_1 = n - p)$ ,  $p = k + 1$ , i.e., compute a p-value. Note, for a two-sided alternative, this will give the same result as using the above  $F$  test.

### 6.7.7 $t$ tests for Omitting a Single $x_j$ or Setting $\beta_j = 0$

For the more particular case of inferring about a single regression function parameter,  $\beta_j$ , (Result 6.10), we have

$$H_0 : \beta_j = b_0 \quad \text{null hypothesis}$$

$$H_1 : \beta_j \neq b_0 \quad \text{alternative hypothesis (or } > \text{ or } <\text{)}.$$

And, we may test the hypothesis via a  $t$  statistic under the null hypothesis,

$$t = \frac{\hat{\beta}_j - b_0}{\sqrt{\hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})_{(jj)}^{-1}}} \sim t(df = n - p).$$

### 6.7.8 $t$ -based Confidence Intervals for Scalar $\mathbf{C}\beta$

Typically, we infer about confidence regions for scalar quantities using  $t$  Results 6.9 or 6.10, which gives regions in the form of intervals.

**Result 6.13** (Interval for General Scalar  $\mathbf{C}\beta$ ). *In the case of Result 6.9, we have*

$$\begin{aligned} \mathbf{C}\hat{\beta} &\pm t(1 - \alpha/2, n - p) \times \widehat{SE}(\mathbf{C}\hat{\beta}) \quad \text{or,} \\ \mathbf{C}\hat{\beta} &- t(1 - \alpha, n - p) \times \widehat{SE}(\mathbf{C}\hat{\beta}), \quad \text{lower bound, or,} \\ \mathbf{C}\hat{\beta} &+ t(1 - \alpha, n - p) \times \widehat{SE}(\mathbf{C}\hat{\beta}), \quad \text{upper bound.} \end{aligned}$$

**Result 6.14** (Interval for the Mean  $E(Y | \mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$ ).

- An important special case of the above Result 6.13 (and 6.7.6) is when  $\mathbf{C} = \mathbf{x}^T$ , some chosen vector of covariates, not necessarily among the observed covariates  $\mathbf{x}_i$ ,  $i = 1, \dots, n$ .
- That is, in this special case, we are inferring about  $E(Y | \mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$ , i.e., about the mean of  $Y | \mathbf{x}$  at some chosen covariate value  $\mathbf{x}$ .

**Result 6.15** (Interval for  $\beta_j$ ).

Specializing the above Result 6.13 even further to a single parameter,  $\beta_j$  (Result 6.10), we have

$$\begin{aligned} \hat{\beta}_j &\pm t(1 - \alpha/2, n - p) \times \widehat{SE}(\hat{\beta}_j) \quad \text{or,} \\ \hat{\beta}_j &- t(1 - \alpha, n - p) \times \widehat{SE}(\hat{\beta}_j), \quad \text{lower bound, or,} \\ \hat{\beta}_j &+ t(1 - \alpha, n - p) \times \widehat{SE}(\hat{\beta}_j), \quad \text{upper bound.} \end{aligned}$$

### 6.7.9 $t$ -based Prediction Intervals for $Y | \mathbf{x}$

In Result 6.14, we *estimated* the mean of a population of values represented by  $Y | \mathbf{x}$ , i.e., we gave interval estimators for  $E(Y | \mathbf{x})$ . Sometimes, we don't want an interval for the mean, but for the random variable  $Y | \mathbf{x}$  itself. Intuitively, given that  $Y | \mathbf{x}$  is centered at  $E(Y | \mathbf{x})$ , but with extra variability according to our model,

$$Y | \mathbf{x} = E(Y | \mathbf{x}) + \epsilon,$$

this next result on **prediction intervals** should be somewhat intuitive. In particular, our **predictor** (BLUP) is

$$\text{Pred}(Y | \mathbf{x}) = \hat{\mu}(Y | \mathbf{x}) = \mathbf{x}^T \hat{\boldsymbol{\beta}}$$

which is the same as the **estimator** of  $E(Y | \mathbf{x})$  as mentioned in Result 6.14. But, now, the **prediction error variance** is

$$\begin{aligned} \text{Var}(\text{Pred}(Y | \mathbf{x}) - Y | \mathbf{x}) &= \text{Var}(\hat{\mu}(Y | \mathbf{x})) + \text{Var}(Y | \mathbf{x}) \\ &= \text{Var}(\mathbf{x}^T \hat{\boldsymbol{\beta}}) + \text{Var}(Y | \mathbf{x}) \\ &= \sigma^2 \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x} + \sigma^2 \\ &= \sigma^2 (1 + \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}), \end{aligned}$$

which leads to the **estimated standard error of prediction**

$$\widehat{\text{SE}}(\text{Pred}(Y | \mathbf{x}) - Y | \mathbf{x}) = \sqrt{MSE(1 + \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x})}$$

and to the associated  $t$  distributional result, similar to what has been given previously,

$$\frac{(\text{Pred}(Y | \mathbf{x}) - Y | \mathbf{x}) - 0}{\widehat{\text{SE}}(\text{Pred}(Y | \mathbf{x}) - Y | \mathbf{x})} \sim t(n - p),$$

which leads to the **prediction interval**

$$\boxed{\text{Pred}(Y | \mathbf{x}) \pm t(1 - \alpha/2, n - p) \widehat{\text{SE}}(\text{Pred}(Y | \mathbf{x}) - Y | \mathbf{x})}.$$

More on prediction, discussed in this section, and on estimation of the mean, in Result 6.14 of the previous section, is given in

***Additional Reading:***

[RS13, 7.4.2, 7.4.3] for SLR

[RS13, 10.2.3, 10.2.4] for MLR

[KNNL05, 2.4, 2.5] for SLR

[KNNL05, 6.7] for MLR

$\mathcal{R}$

[KNNL05] also discuss simultaneous inference for means and for predictions as well as predicting the average of  $m$  values of  $Y|\mathbf{x}$  ([KNNL05, 4.2, 4.3, 6.7]). See [RS13, pg 190 & Sec. 7.4.3] and [RS13, pg 283-4 & Sec. 10.4.2] for similar discussions of these procedures in the SLR and MLR cases, respectively. These can be subsumed into our  $\mathbf{C}\boldsymbol{\beta}$  framework, and we skip the details.

## 6.8 Example Data Analysis Using $t$ and $F$ Results

We use a small, simple (but interesting) data set to illustrate the practical essence of the results just presented, above. We use expressions of the above sections in “by-hand” computations (in  $\mathbf{R}$ ), because we are in Dr. Barber’s class, and because we want to see how our results are related to more high-level functions in  $\mathbf{R}$ , which hide such underlying details. Once we understand the underlying details, we would typically use the high-level  $\mathbf{R}$  functions in practice, of course.

The data are measurements of the distance between Earth and  $n = 24$  nebulae ( $y$ ) and the velocity with which these nebulae are traveling from Earth ( $x = \text{velocity}$ , or  $\mathbf{x}^T = (1, \text{velocity})$  in our simple linear regression model example, here). I call these data the **big bang data**. I will explain these data in class, and have more discussion throughout the analysis. See [RS13, Ch. 7] and the related  $\mathbf{R}$  library package **Sleuth3** for more information.

To be sure, we compute various results in previous sections associated with the (simple) linear regression function,  $E(Y | x) = \beta_0 + \beta_1 x$ .

The next chunk simply reads the data into a data frame and gives a quick look at it—check for gross errors.

```
> ## LS estimation of beta
> ## File is called case0701 in Sleuth3 package
> bigbang.df<- Sleuth3::case0701
> names(bigbang.df)

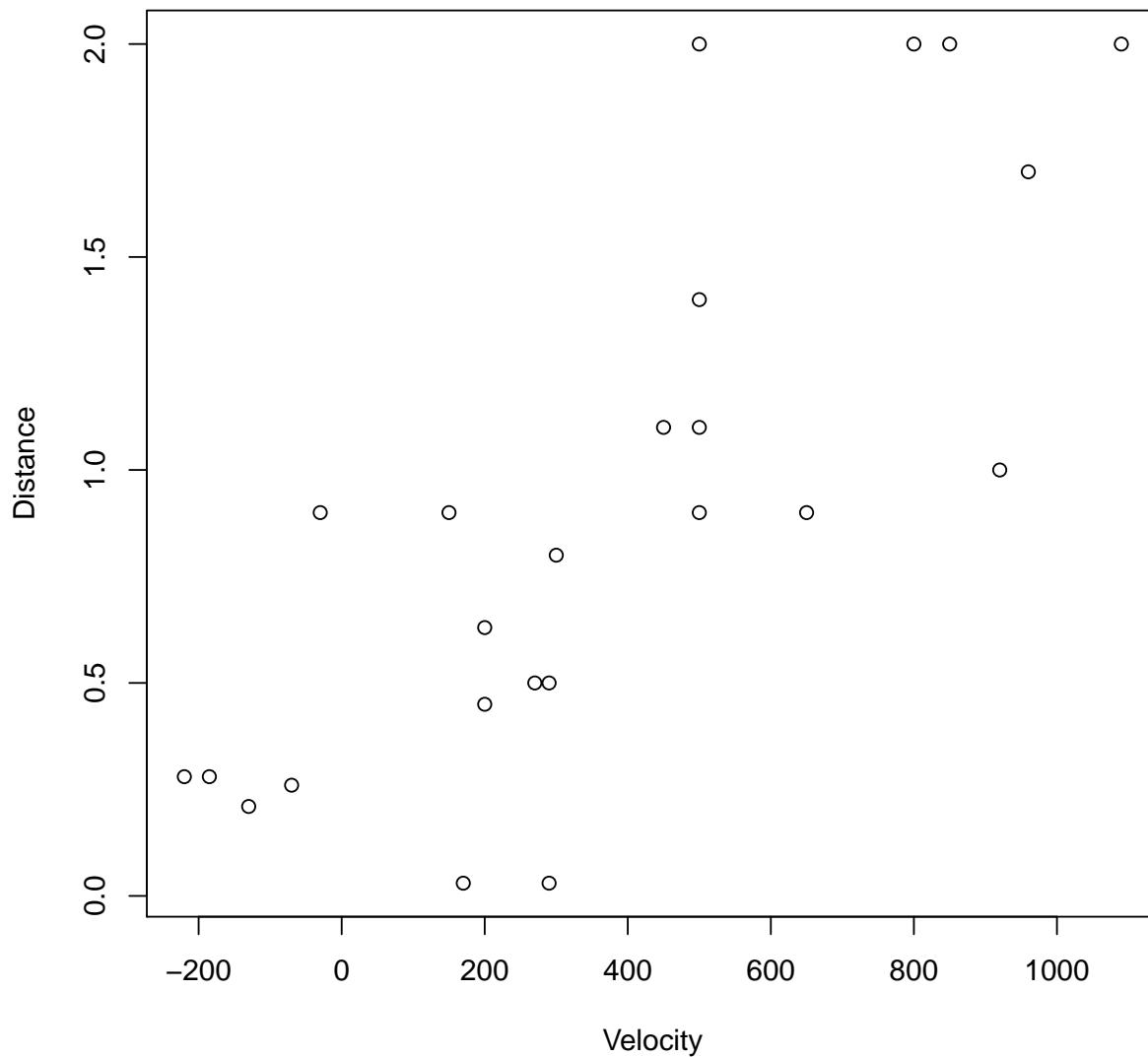
[1] "Velocity" "Distance"

> head(bigbang.df)

  Velocity Distance
1      170     0.03
2      290     0.03
3     -130     0.21
4      -70     0.26
5     -185     0.28
6     -220     0.28
```

Quick graphical EDA in the next chunk—check for gross errors.

```
> plot(Distance ~ Velocity, data=bigbang.df)
```



- The next chunk prepares model components mentioned in above type-set notes. I try to use suggestive object names!
- Notice how the matrix-vector code in the following and subsequent

code would work not just for SLR, but for any normal linear model, as we said of the above  $t$  and  $F$  distributional results.

```
> y<- bigbang.df$Distance
> X<- model.matrix(Distance ~ Velocity, data=bigbang.df)
> head(y,n=3); tail(y,n=3)

[1] 0.03 0.03 0.21
[1] 2 2 2

> (n<-dim(X)[1]); (p<- dim(X)[2]); head(X,n=3); tail(X,n=3)

[1] 24
[1] 2
  (Intercept) Velocity
1           1      170
2           1      290
3           1     -130
  (Intercept) Velocity
22          1      850
23          1      800
24          1     1090
```

Obtain quantities discussed in above typeset notes, first by hand.

```
> (betahat<- (XtXinv<- solve( XtX<-t(X) *% X )) *% t(X) *% y)

[,1]
(Intercept) 0.399170440
Velocity    0.001372408

> ehat<- (y - X *% betahat)
> ## MSE (estimator of sigma^2)
> (sig2hat<- as.vector(t(ehat) *% ehat / (n - p)))

[1] 0.1645359
```

```
> sqrt(sig2hat)
[1] 0.4056302

> varBhat<- sig2hat * XtXinv
> (seBhat<- sqrt(diag(varBhat)))

(Intercept)      Velocity
0.1186661507  0.0002278214
```

Now, we compute more automatically, as if you were not taking Dr. Barber's class! In class, we will point out the correspondence of the output, below, with our "by hand" results, above.

```
> ## More automatically
> summary(bigbang.lm<- lm(Distance ~ Velocity, data=bigbang.df))

Call:
lm(formula = Distance ~ Velocity, data = bigbang.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.76717 -0.23517 -0.01083  0.21081  0.91463 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.3991704  0.1186662   3.364   0.0028 **  
Velocity    0.0013724  0.0002278   6.024 4.61e-06 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.4056 on 22 degrees of freedom
Multiple R-squared:  0.6226, Adjusted R-squared:  0.6054 
F-statistic: 36.29 on 1 and 22 DF,  p-value: 4.608e-06
```

*t*-based confidence interval estimation of **slope** using “ $\mathbf{C}\boldsymbol{\beta}$ ” form, as discussed in above typeset notes.

```
> ## slope interval
> (Cmat<- matrix(c(0,1),nrow=1,ncol=p))

[,1] [,2]
[1,]    0    1

> (CBhat<- as.vector(Cmat%*%beta))
[1] 0.001372408

> (CBse<- as.vector(sqrt(Cmat %*% varBhat %*% t(Cmat))))
[1] 0.0002278214

> (tmult<- qt(1-0.05/2, df=n-p))
[1] 2.073873

> CBhat + c(-1,1) * CBse * tmult
[1] 0.0008999349 0.0018448801
```

```
> ## slope interval more automatically
> confint(bigbang.lm, parm=2)

              2.5 %      97.5 %
Velocity 0.0008999349 0.00184488
```

*t*-based test of **intercept** using the “ $\mathbf{C}\boldsymbol{\beta}$ ” approach, as discussed in above typeset notes. In class, we will compare these “by hand” computation with what we already obtained more automatically, above.

```

> ## t-test intercept
> (Cmat<- matrix(c(1,0),nrow=1,ncol=p))

      [,1] [,2]
[1,]    1    0

> (CBhat<- as.vector(Cmat%*%beta))
[1] 0.3991704

> (CBse<- as.vector(sqrt(CBvar<- Cmat %*% varBeta %*% t(Cmat))))
[1] 0.1186662

> b0<- 0
> (tstat<- (CBhat - b0) / CBse)
[1] 3.36381

> (pval<- 2 * pt(abs(tstat), df=n-p, lower=FALSE))
[1] 0.002803008

> (reject <- pval <= 0.05)
[1] TRUE

> ## equivalent (if unconventional) F-test
> (Fstat<- as.vector(t(CBhat - b0) %*% solve(CBvar) %*%
+ (Cmat%*%beta - b0)))
[1] 11.31522

> (pval<- pf(Fstat, 1, n-p, lower.tail=FALSE))
[1] 0.002803008

> tstat^2 == Fstat
[1] TRUE

> ## test intercept more automatically: see summary results, above

```

$t$ -based confidence interval estimation of **mean** via “ $\mathbf{C}\beta$ ”, as discussed in above typeset notes. That is, we want to estimate  $E(Y|x) = \beta_0 + \beta_1 x$  at some value of  $x$ . For illustration, we take  $x = 800$ , so that  $\mathbf{C} = [1, 800]$ , a row matrix. First, by hand.

```
> ## Confidence interval for mean,  $E(y) = mu(x, beta)$ , at  $x = (1, 800)^T$ 
> (Cmat<- matrix(c(1,800),nrow=1,ncol=p))

      [,1] [,2]
[1,]    1   800

> (CBhat<- as.vector(Cmat%*%beta))

[1] 1.497096

> (CBse<- as.vector(sqrt(Cmat %*% varBhat %*% t(Cmat)))) 

[1] 0.1277242

> (tmult<- qt(1-0.05/2, df=n-p))

[1] 2.073873

> CBhat + c(-1,1) * CBse * tmult

[1] 1.232213 1.761980
```

Then, more automatically, as in practice, after being armed with the confidence of having taking Dr. Barber's class.

```
> ## Confidence interval more automatically
> xnew<- data.frame(Velocity=800)
> predict(bigbang.lm, newdata=xnew, se.fit=TRUE, interval="confidence",
+           level=0.95)

$fit
      fit      lwr      upr
```

```

1 1.497096 1.232213 1.76198

$se.fit
[1] 0.1277242

$df
[1] 22

$residual.scale
[1] 0.4056302

```

*t*-based prediction interval estimation of  $y$  using the “ $\mathbf{C}\boldsymbol{\beta}$ ” approach, as discussed in above typeset notes.

```

> ## Prediction interval for new  $y = \mu(x, \beta) + \epsilon$  at  $x = (1, 800)^T$ 
> (Cmat<- matrix(c(1,800),nrow=1,ncol=p))

[,1] [,2]
[1,]    1   800

> (CBhat<- as.vector(Cmat%*%betahat))

[1] 1.497096

> (CBse<- as.vector(sqrt(sig2hat + Cmat %*% varBhat %*% t(Cmat)))))

[1] 0.4252638

> (tmult<- qt(1-0.05/2, df=n-p))

[1] 2.073873

> CBhat + c(-1,1) * CBse * tmult

[1] 0.6151532 2.3790397

> ## Prediction interval more automatically
> predict(bigbang.lm, newdata=xnew, se.fit=TRUE, interval="predict",
+           level=0.95)

```

```
$fit
  fit      lwr      upr
1 1.497096 0.6151532 2.37904

$se.fit
[1] 0.1277242

$df
[1] 22

$residual.scale
[1] 0.4056302
```

We did not see an implementation of the “F v R” approach (§6.7.4) in this example, but will see it using the `anova` function in the next chapter.

## 6.9 Summary and Final Remarks

Despite what may appear to be several different methods for estimation and testing given above in §6.7, in essence, everything we have done is as summarized in §6.7.1, which are based on only a few distributional results in §6.6.3 that follow from the assumptions of our normal linear model.

While we did not discuss formally in this chapter of our notes the long-run relative frequency interpretation of the above frequentist tests and intervals, this interpretation remains. (Hopefully, I said something about this in class.) Remember, in lecture note chapter 1, we said that we often use mathematical models (normal and the related distributions seen here) to approximate randomization or sampling distributions, wherein the notion of **hypothetical replications** and the frequentists’ **long-run relative frequency** interpretation are more easily seen.

In the next chapter, we will give a relatively full analysis of a slightly more interesting data set. There, we will introduce other high level functions in `R` to conduct the tests/intervals that we have introduced here. In particular, I find the functions `glh.test` and `estimable` functions in the

R library package, `gmodels`, to be useful for the sort of “ $\mathbf{C}\boldsymbol{\beta}$ ” inference approach. Will will continue to illustrate the “F vs. R” model approach (aka extra sum-of squares approach), too, in case you prefer that equivalent approach to testing. Also, we will continue to introduce additional concepts, such as the overall  $F$ -test and  $R^2$ , etc.

# Lecture 7

## Linear Models III: Example Frequentist Data Analysis

### Contents

---

|             |  |            |
|-------------|--|------------|
| <b>7.1</b>  | <b>Introduction</b>  | <b>219</b> |
| 7.1.1       | Overview   | 219        |
| 7.1.2       | Preview  | 221        |
| <b>7.2</b>  | <b>Data Set</b>  | <b>222</b> |
| <b>7.3</b>  | <b>Graphical Exploratory Data Analysis (EDA)</b>               | <b>225</b> |
| <b>7.4</b>  | <b>Initial Linear Regression Model</b>                         | <b>233</b> |
| <b>7.5</b>  | <b>Omitted &amp; Added Variable Plots</b>                      | <b>234</b> |
| <b>7.6</b>  | <b>Remodeling</b>  | <b>237</b> |
| <b>7.7</b>  | <b>Observed vs. Fitted Plot</b>                                | <b>238</b> |
| <b>7.8</b>  | <b>Residual Plots</b>  | <b>239</b> |
| <b>7.9</b>  | <b>Overall F Test: Special Case of <math>C\beta</math></b>     | <b>242</b> |
| 7.9.1       | anova Function for F v R Approach                              | 244        |
| 7.9.2       | The glh.test R Function  | 247        |
| <b>7.10</b> | <b><math>R^2</math> &amp; Adjusted <math>R^2</math></b>        | <b>248</b> |
| <b>7.11</b> | <b>Using an Interaction Term</b>                               | <b>250</b> |
| <b>7.12</b> | <b><math>t</math>-based inference for <math>\beta_j</math></b> | <b>257</b> |
| 7.12.1      | Default lm Printout and summary                                | 257        |
| 7.12.2      | By Hand Test and Intervals                                     | 258        |
| 7.12.3      | The confint Function   | 259        |
| 7.12.4      | The estimable Function   | 259        |
| <b>7.13</b> | <b>Qualitative Covariates</b>                                  | <b>260</b> |

|   |            |
|---|------------|
| 7.13.1 Cell Reference Coding . . . . .  | 260        |
| <b>7.14 Intervals for <math>E(Y   \mathbf{x})</math> &amp; <math>Y   \mathbf{x}</math> with predict . . . . .</b> | <b>266</b> |
| 7.14.1 Another Warning About Extrapolation . . . . .  | 268        |
| <b>7.15 Summary . . . . .</b>   | <b>269</b> |

---

***Main Objectives:***

- Be aware of problems stemming from multiple comparisons (tests & intervals) in the exploratory setting.
- Always report as precisely as possible the procedure/process taken to arrive at a final inference(s) or statistical models. If you p-hack, say so.
- General linear hypothesis
- Estimating general linear combinations  $\mathbf{C}\boldsymbol{\beta}$
- Full vs reduced (F v R) approach (aka extra sum of squares approach)
- Coefficient of determination  $R^2$ . Adjusted  $R^2$ .
- `lm`, `anova`, `confint`, `predict`, `glh.test`, `estimable`
- Plotting (various types and R functions)
- Cell reference coding (aka treatment or corner-point coding) for a qualitative (factor) variable

---

$\mathcal{O}$

***Additional Reading:***

- [Wak13, §4.5, 4.7, 5.6, 5.11, 5.13]
- [KNNL05, Chap. 1,2,6-8]
- [RS13, Chap. 7, 9 & 10]

---

 $\mathcal{R}$ 

## 7.1 Introduction

### 7.1.1 Overview

Here, we touch on several aspects of a reasonably thorough analysis of data from a randomized experiment, to be discussed, shortly.

- Our analysis is presented in a somewhat **exploratory** manner in the sense that we have not decided on a model before we look at the data; we have not decided on a model or set of hypotheses or set of inferences *a priori* ([Wak13, §4.5, 4.7]).
- An exploratory analysis lets the data guide the selection of models and inferences—*a posteriori*, after seeing the data. This is potentially misleading (and potentially dangerous depending on the context) because, if we let the data suggest a series of models before selecting a final model, then typical **frequentist properties are difficult to justify**.
- In particular, actual **error rates**, **p-values** and interval **coverage rates** likely will differ from the nominal values that you might report.
- Generally speaking the exploratory process must be incorporated into the inference framework so that the process can be accounted for in

the properties of the resulting inferences. This is usually difficult to do in such exploratory analyses.

- On the other hand, in the case that we specify *a priori* a **fixed number of tests/inferences**, methods exist to incorporate this sort of **pre-specified multiple comparison process** (multiple inference) so that error rates and coverage rates are somehow controlled ([Wak13, §4.6]).
- Several **classic procedures** allow such control when multiple inferences are specified *a priori*. These so-called multiple comparison procedures go by the names **Tukey(-Kramer)** (**honestly significant difference (HSD)**), **Scheffé**, **Bonferroni**, **Working-Hotelling**, and depend on the kind of inferences being conducted; see, e.g., the index of [KNNL05]. [Wak13, §4.6] also discusses Bonferroni adjustment of error rates in addition to (positive) **false discovery rates** ((p)FDR), including related Bayesian approaches.
- See the short discussion by [Wak13, §4.7] of the three scenarios: confirmatory, exploratory, prediction.

- We will not likely have time to return to a more systematic and thorough discussion of the multiple comparison problem in this class. Just be aware of the problems mentioned here.
- In any case, we should always follow good (honest) practice of reporting the exact procedure(s) taken to arrive at a final model or inference.
- As an example of what **not** to do: tell your student to go find interesting results in a data set, using any of the methods they know or

can learn. Then, reverse engineer an interesting question/hypothesis along to fit your final result so that it appears as if your result confirms an a priori (set of) inference(s) while hiding the above mentioned problems associated with an exploratory approach.

See the popular article on <https://www.vox.com/science-and-health/2018/9/19/17879102/brian-wansink-cornell-food-brand-lab-retractions-jama>.

- A very closely related treatment of variable selection and its effect on bias and variance of inference procedures (so-called **bias-variance trade-off**) is typically discussed in a class like INF 504.

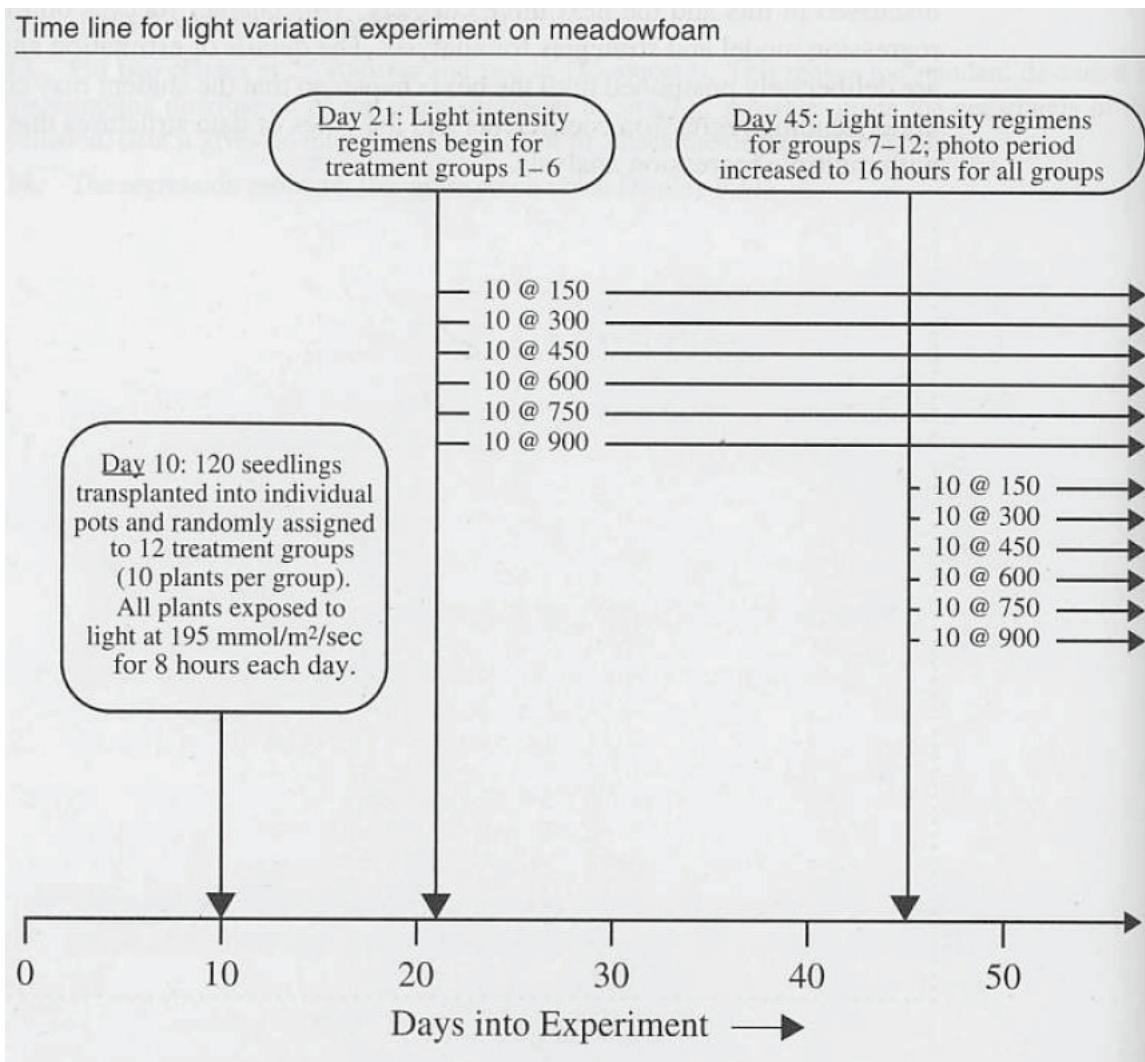
### 7.1.2 Preview

- In a very real sense, our analysis here simply follows the  $t$  and  $F$  results introduced in chapter 6 of our notes. You should study the example here and the Big Bang example of §6.8 until you understand the common structure of these analyses, which, ultimately, can be reduced to inference about some linear combinations,  $\mathbf{C}\boldsymbol{\beta}$ . In a subsequent chapter, when discussing ANOVA, we will again use these results, essentially unchanged.
- In the course of our analysis, we introduce the “**overall F test**,” which is just a special case of §6.7.3 or 6.7.4 using the  $F$  distribution Result 6.11.
- We introduce the **coefficient of determination**,  $R^2$ , a commonly used measure of the linear association of the response with the covariates.
- And, we introduce the **coding of a categorical covariate**, which we will revisit, later, with ANOVA.

- We will make an attempt to illustrate computations in a few equivalent ways. In practice, you will likely choose the more automated ways provided by R, though I hope our more “by hand” approaches will help to illuminate these more high-level ways, which tend to hide detail for conciseness.

## 7.2 Data Set

**Example 7.1** (Meadowfoam Seed Oil Experiment). *Meadowfoam (*Limnanthes alba*)* ([RS13, Study 9.1.1]) is a plant valued for its edible seed oil, with applications to cosmetic and other industries. Interest lies in creating more seeds, hence in more flowers on a plant. We investigate the relationship between the number of flowers on a plant,  $Y$ , and two covariates. The first covariate  $X_1$  is the intensity of light ( $\mu\text{mol}/\text{m}^2/\text{sec}$ ) under which plants are grown, the “light regimen”; there are six unique levels. The second covariate  $X_2$  is the time at which the light regimen was started in days before photoperiod floral induction (PFI); two unique levels. Ten seedlings were randomly assigned to each of the  $6 \times 2 = 12$  treatment levels for a total of  $n = 120$  seedlings. The entire experiment was replicated (in 2 blocks, which we ignore) for a total of  $n = 240$ . NOTE HOWEVER: Our data consist of averages of flower counts for 10 plants with 2 groups of 10 plants for each of the 12 treatment combinations ( $n = 24$  for us). We ignore this, too. See figures that follow.



Average number of flowers per meadowfoam plant, in 12 treatment groups

|        | Light Intensity ( $\mu\text{mol/m}^2/\text{sec}$ ) |      |      |      |      |      |
|--------|--|------|------|------|------|------|
|        | 150  | 300  | 450  | 600  | 750  | 900  |
| Timing | 62.3   | 55.3 | 49.6 | 39.4 | 31.3 | 36.8 |
|        | 77.4   | 54.2 | 61.9 | 45.7 | 44.9 | 41.9 |
| Timing | 77.8   | 69.1 | 57.0 | 62.9 | 60.3 | 52.6 |
|        | 75.6   | 78.0 | 71.1 | 52.2 | 45.6 | 44.4 |

The following chunk reads the data from a file. Notice we create a qualitative (non-quantitative) variable (factor) with meaningful labels. We will return to this when discussing qualitative (factor) covariates in §7.13, below, and again when discussing ANOVA. See [Wak13, §5.5.2] for a very concise discussion of coding factors and [RS13, Sec. 9.3], [KNNL05, pp. 218-219] and [KNNL05, Chap.8] for more discussion.

```
> ### Meadowfoam (Limnanthes alba) data from Statistical Sleuth 3e Case
> ### Study 9.1.1.
>
> flower.df<- Sleuth3::case0901
> head(flower.df,n=3);tail(flower.df,n=3)

  Flowers Time Intensity
1     62.3     1       150
2     77.4     1       150
3     55.3     1       300
  Flowers Time Intensity
22    45.6     2       750
23    52.6     2       900
24    44.4     2       900

> ### R is currently seeing every variable (i.e., data frame column,
> ### i.e., list component) as "numeric" (or "integer"):
>
> str(flower.df)

'data.frame': 24 obs. of  3 variables:
 $ Flowers : num  62.3 77.4 55.3 54.2 49.6 ...
 $ Time    : int  1 1 1 1 1 1 1 1 1 ...
 $ Intensity: int  150 150 300 300 450 450 600 600 750 750 ...

> ### We create a categorical variable, i.e., factor, from the Time
> ### variable. This will be useful to illustrate the use of a
> ### categorical covariate later. Note we create labels here
> ### so that we get something more meaningful than "1" or "2" for
> ### Time:
>
> flower.df$Time <- factor(flower.df$Time, levels=c(1,2),
```

```

+
+                               labels=c("0 days", "24 days"),
+                               ordered=TRUE)
>
> ### The Time variable really refers to a period of time in days so we
> ### create a quantitative variable corresponding to Time (call it
> ### Days): Time=1 means 0 days prior to PFI, Time=2 means 24 days:
>
> flower.df$Days <- rep(c(0,24),table(flower.df$Time)) ## BE CAREFUL!!
> head(flower.df, n=3); tail(flower.df,n=3)

  Flowers    Time Intensity Days
1   62.3 0 days      150     0
2   77.4 0 days      150     0
3   55.3 0 days      300     0
  Flowers    Time Intensity Days
22  45.6 24 days     750    24
23  52.6 24 days     900    24
24  44.4 24 days     900    24

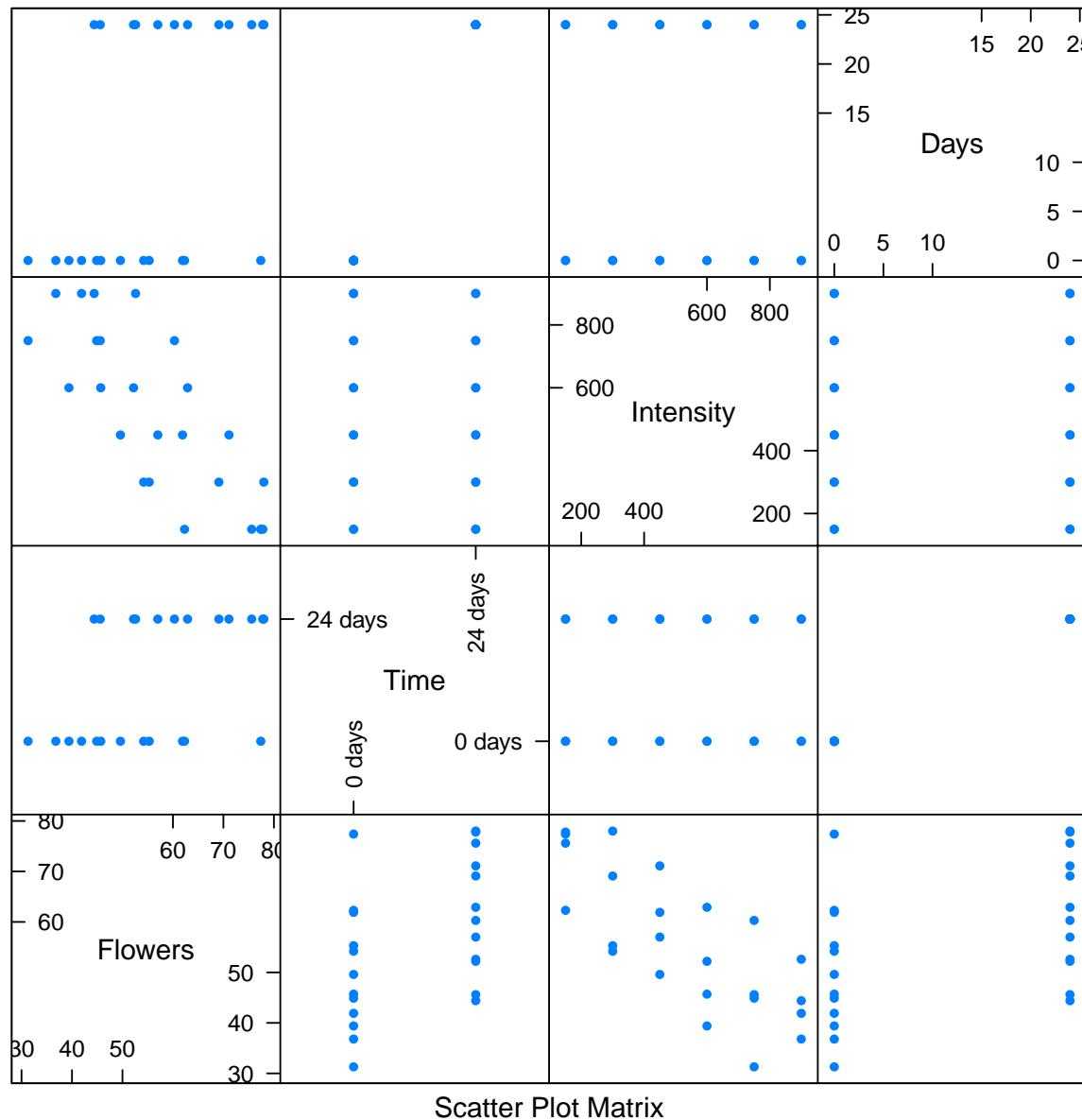
```

Note, the actual analysis of these “flower data” by their original investigators may very well have been more **confirmatory**, with an *a priori* well-defined small set of models/inferences. For example, they may have chosen (*a priori*) to conduct a 2-way ANOVA with a few (pre-specified) comparisons among means—no model/variable selection, no letting data guide inference. After all, it appears that too much effort was made to collect these data, in a controlled experiment, to conduct only exploratory inferences, whose resulting inferences’ properties are difficult to justify. So, our exploratory approach, here, may be a bit contrived, though hopefully instructive.

## 7.3 Graphical Exploratory Data Analysis (EDA)

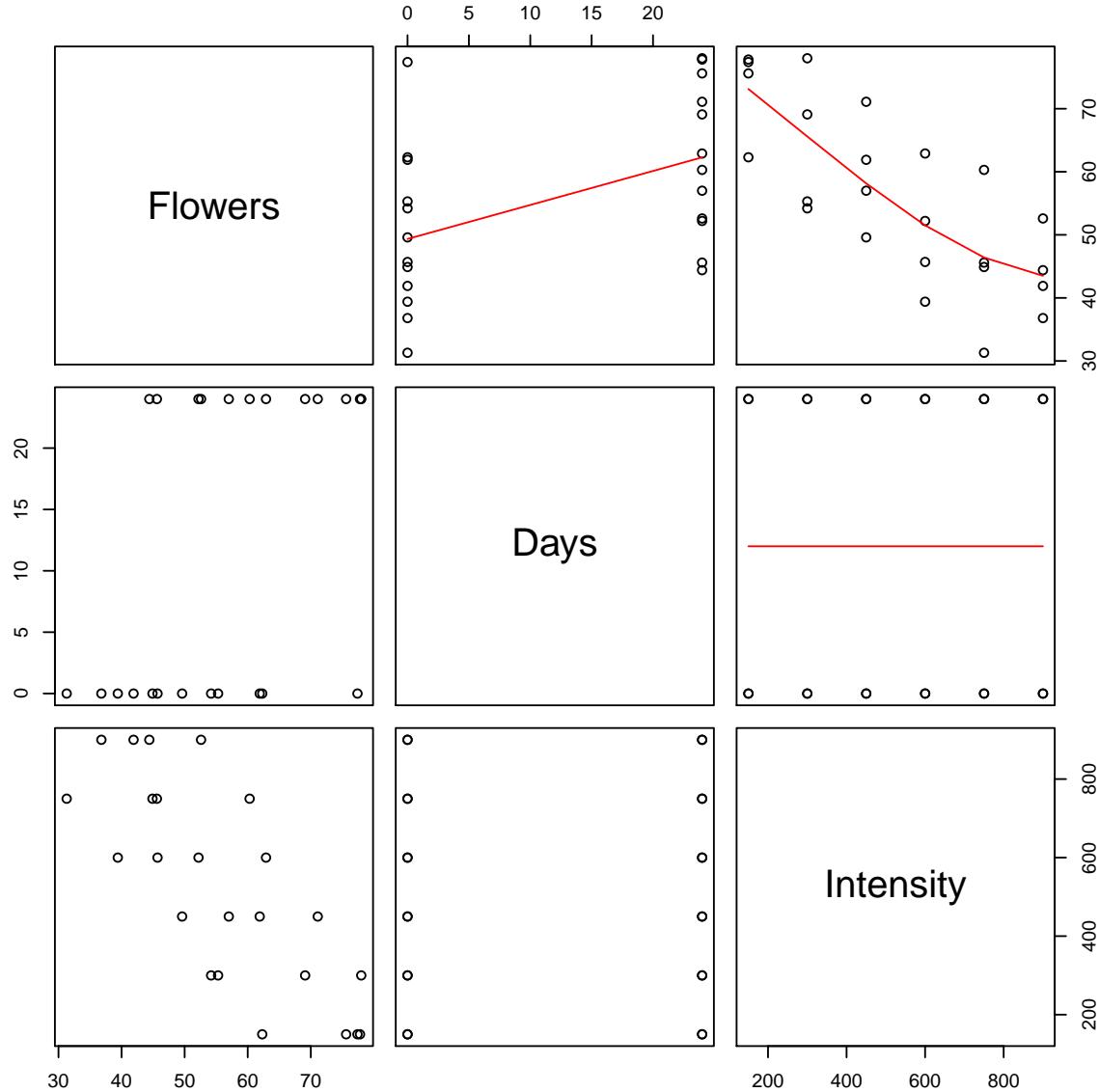
Subsequent Chunks follow with some graphical exploratory data analysis (EDA) to help us formulate regression function models; see [RS13, Sec. 9.5].

```
> ### Some graphical EDA (Exploratory Data Analysis) follows. Some
> ### plots may be preferable to other for certain aspects of the data.
>
> library(lattice)
>
> ### We know what scatter plots and/or dot plots are. How about a
> ### matrix of them (splom for "s"scatter "plo"t "m"atrix): Note that
> ### some matrix elements are not terribly informative:
>
> trellis.par.set("background", list(alpha=1,col="white"))
>
> splom(~flower.df,pch=20)
```

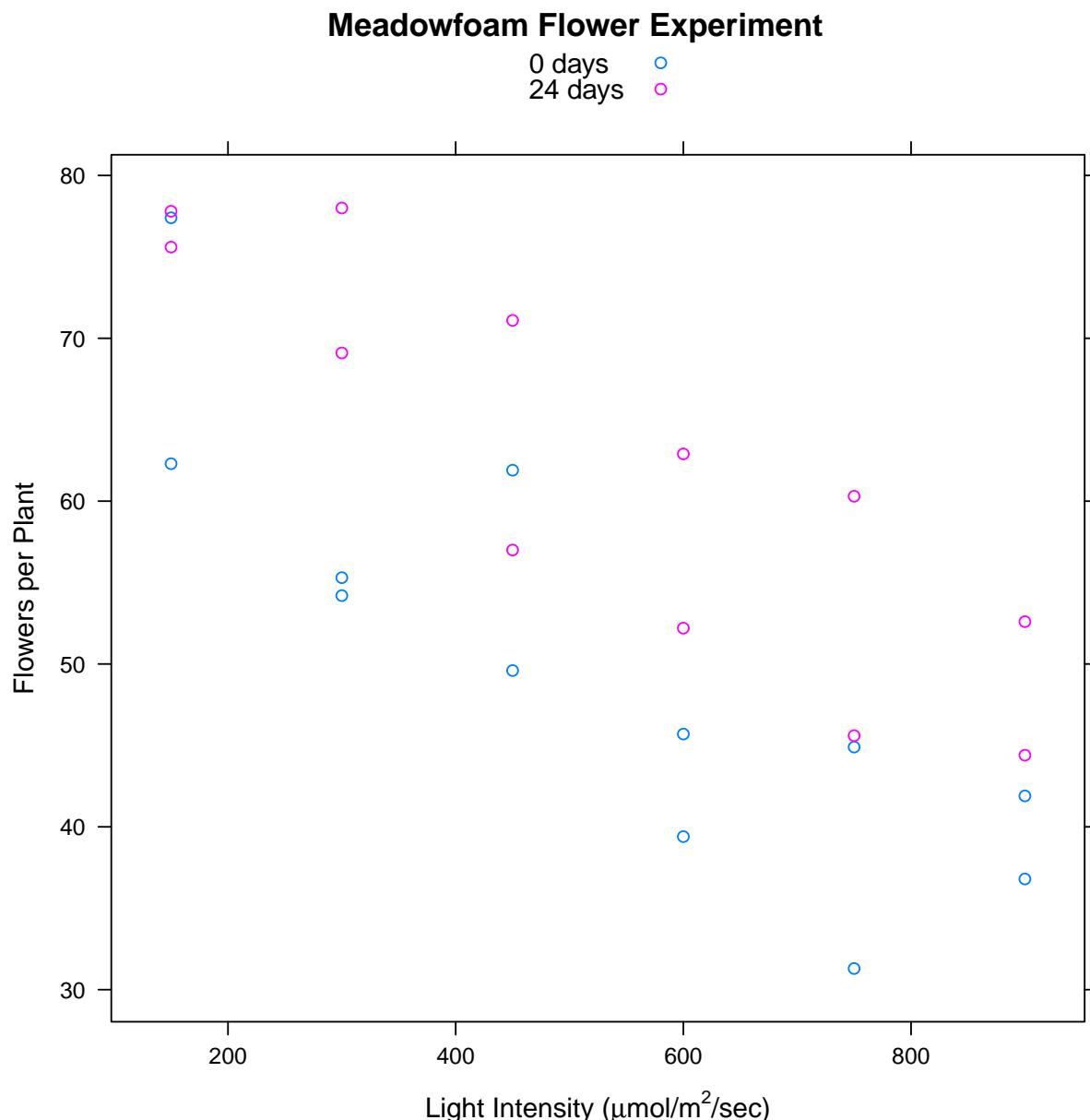


```
> detach(package:lattice)
```

```
> ### Or, similar to previous chunk  
> pairs(Flowers ~ Days + Intensity, data=flower.df, upper.panel=panel.smooth)
```

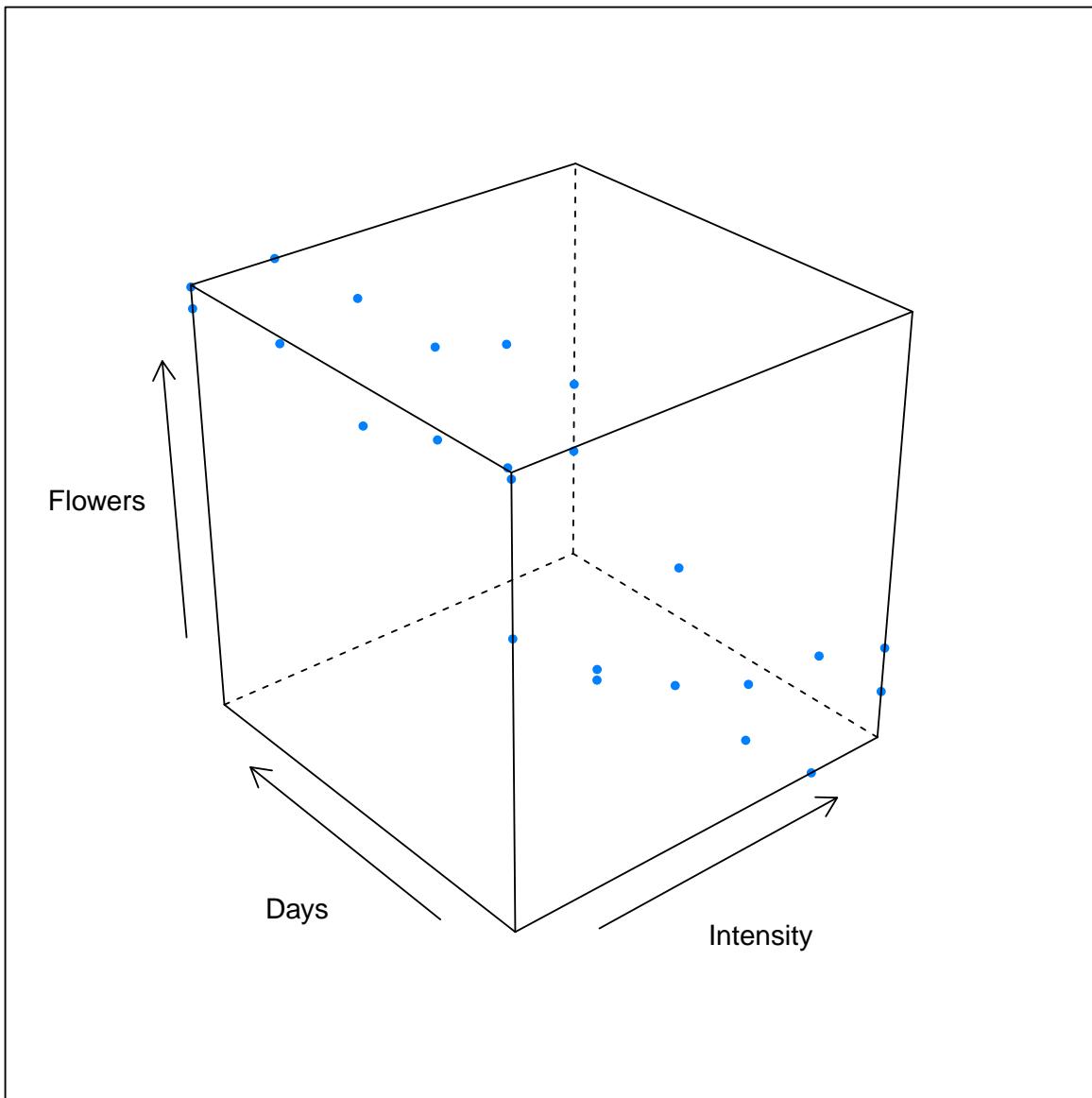


```
> library(lattice)
> ### Or, yet another plot:
> xyplot(Flowers ~ Intensity, data=flower.df, groups=Time,
+         xlab=expression(paste("Light Intensity (",mu,"mol/",m^2,"/sec)", 
+         sep="")),
+         ylab="Flowers per Plant", auto.key=TRUE,
+         main="Meadowfoam Flower Experiment")
```



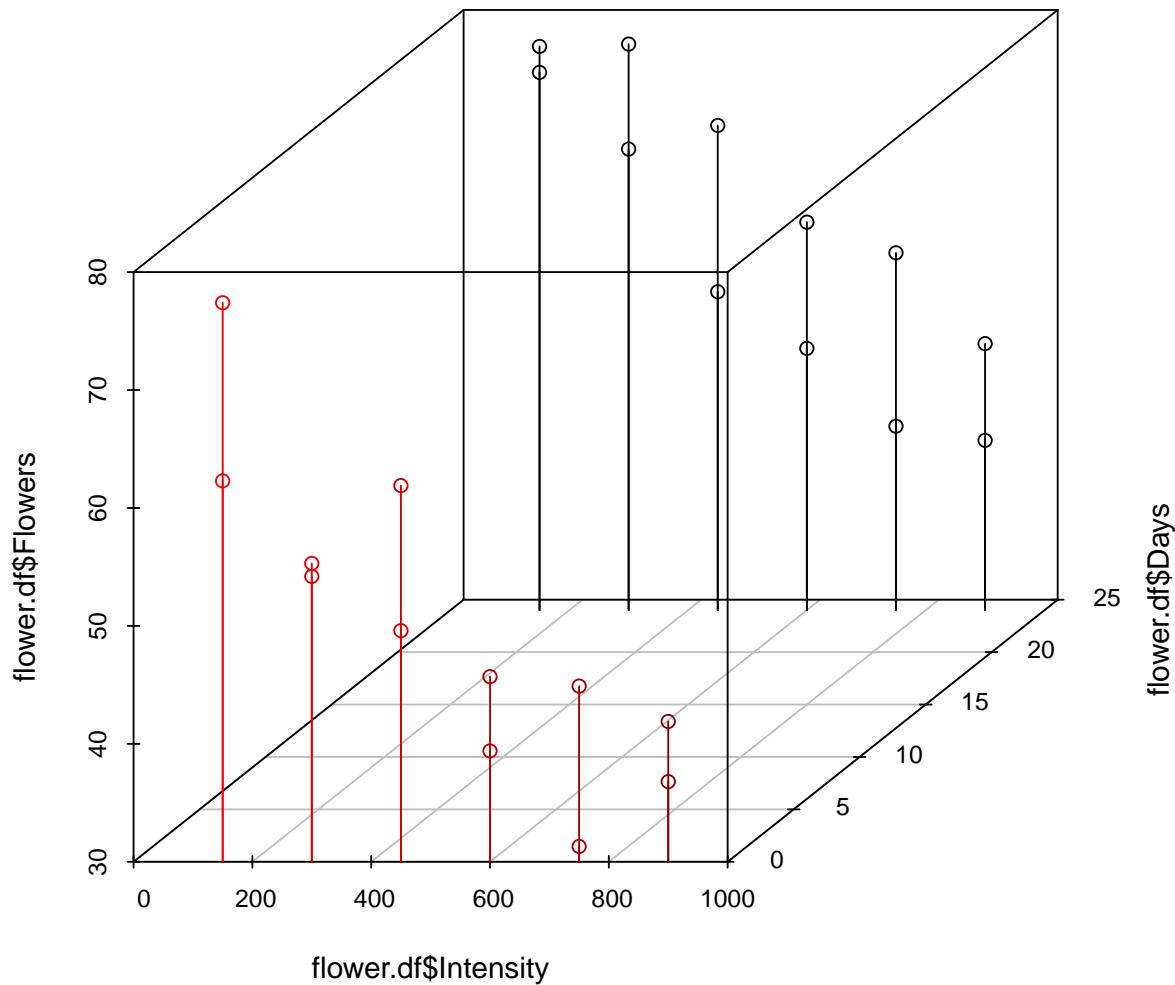
```
> ##detach(package:lattice)
```

```
> ### Next does not seem like such a good plot (the default version  
> ### anyway):  
> library(lattice)  
> cloud(Flowers ~ Intensity * Days, data=flower.df, pch=20)
```



```
> detach(package:lattice)
```

```
> ### Try a similar plot in another library:
> library(scatterplot3d)
> scatterplot3d(x=flower.df$Intensity, y=flower.df$Days,
+                 z=flower.df$Flowers, type=c("h"),
+                 highlight.3d=TRUE)
```



```
> detach(package:scatterplot3d)
>
> ### Enough graphical exploration for now.
```

## 7.4 Initial Linear Regression Model

The above plots suggest that Intensity and Days are candidates for covariate ( $x$ ) variables in an MLR.

### An MLR Model for the Flower Data

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \quad \epsilon_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2)$$

where

- $X_{i1}$  = light intensity for ith plant (Intensity)
- $X_{i2}$  = days before photoperiod floral induction (PFI) for ith plant (Days)

- See [RS13, Display 9.5]; their “time” is our “Days” and their “light” is our “Intensity.”
- How do we write this model in matrix terms?

We begin our modeling with an SLR in the following Chunk, on our way to the MLR.

```
> ### It's obvious from at least one of the above plots that both Intensity
> ### and Days (or Time) are potential predictor variables. But we
> ### start with and SLR on Intensity:
>
> flower1.lm<- lm(Flowers ~ Intensity, data=flower.df)
> summary(flower1.lm)
```

Call:

```
lm(formula = Flowers ~ Intensity, data = flower.df)
```

Residuals:

```

      Min       1Q    Median      3Q      Max
-15.7314 -7.8052   0.0186   6.1857  13.2686

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 77.385000  4.161186 18.597 6.06e-15 ***
Intensity   -0.040471  0.007123 -5.682 1.03e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.94 on 22 degrees of freedom
Multiple R-squared:  0.5947, Adjusted R-squared:  0.5763
F-statistic: 32.28 on 1 and 22 DF,  p-value: 1.03e-05

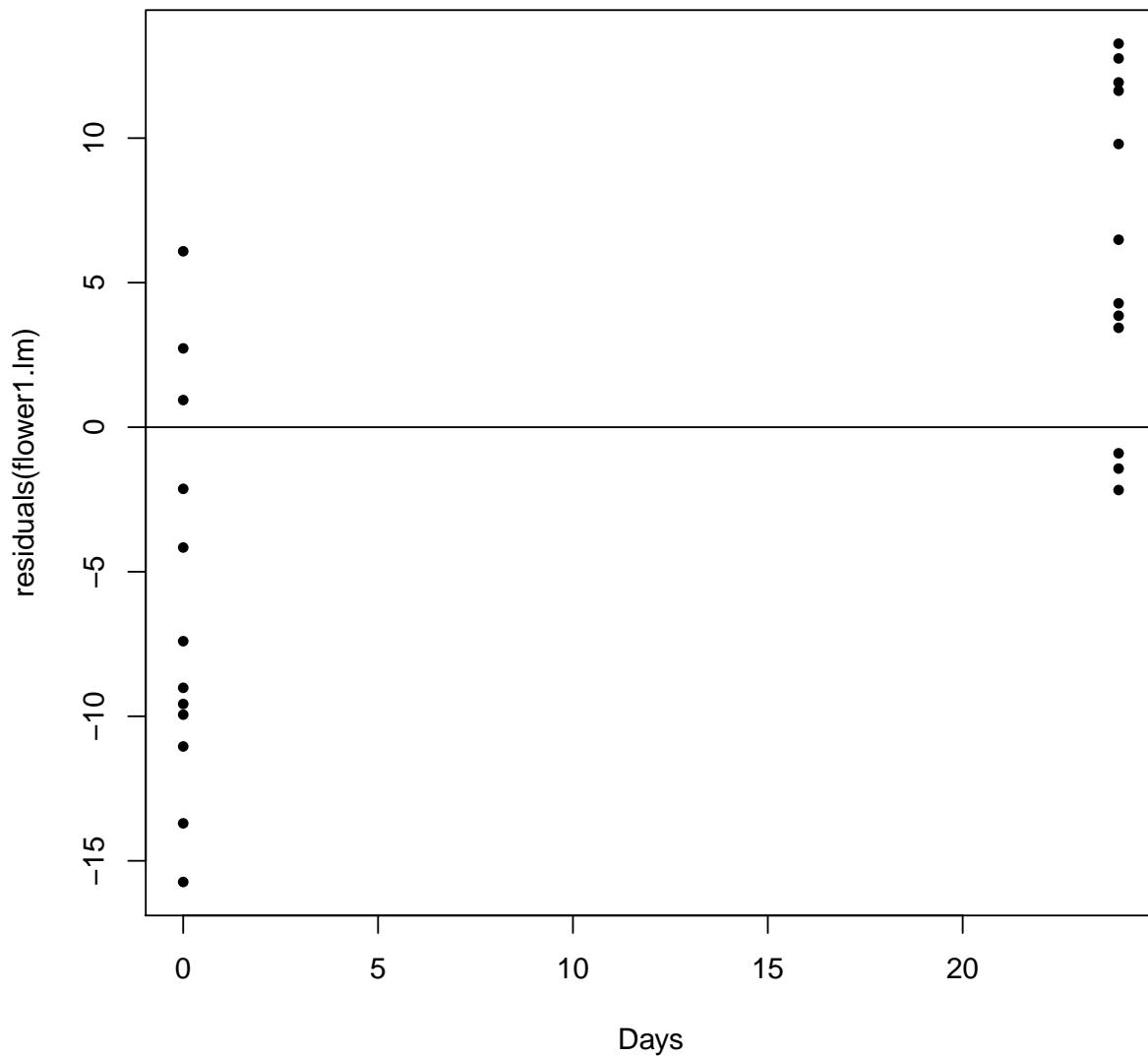
```

- What is the  $X$  matrix for the above SLR?
- What is the estimate for the regression parameter in terms of matrices/vectors?
- Do you recall how to interpret the remainder of the R output?

## 7.5 Omitted & Added Variable Plots

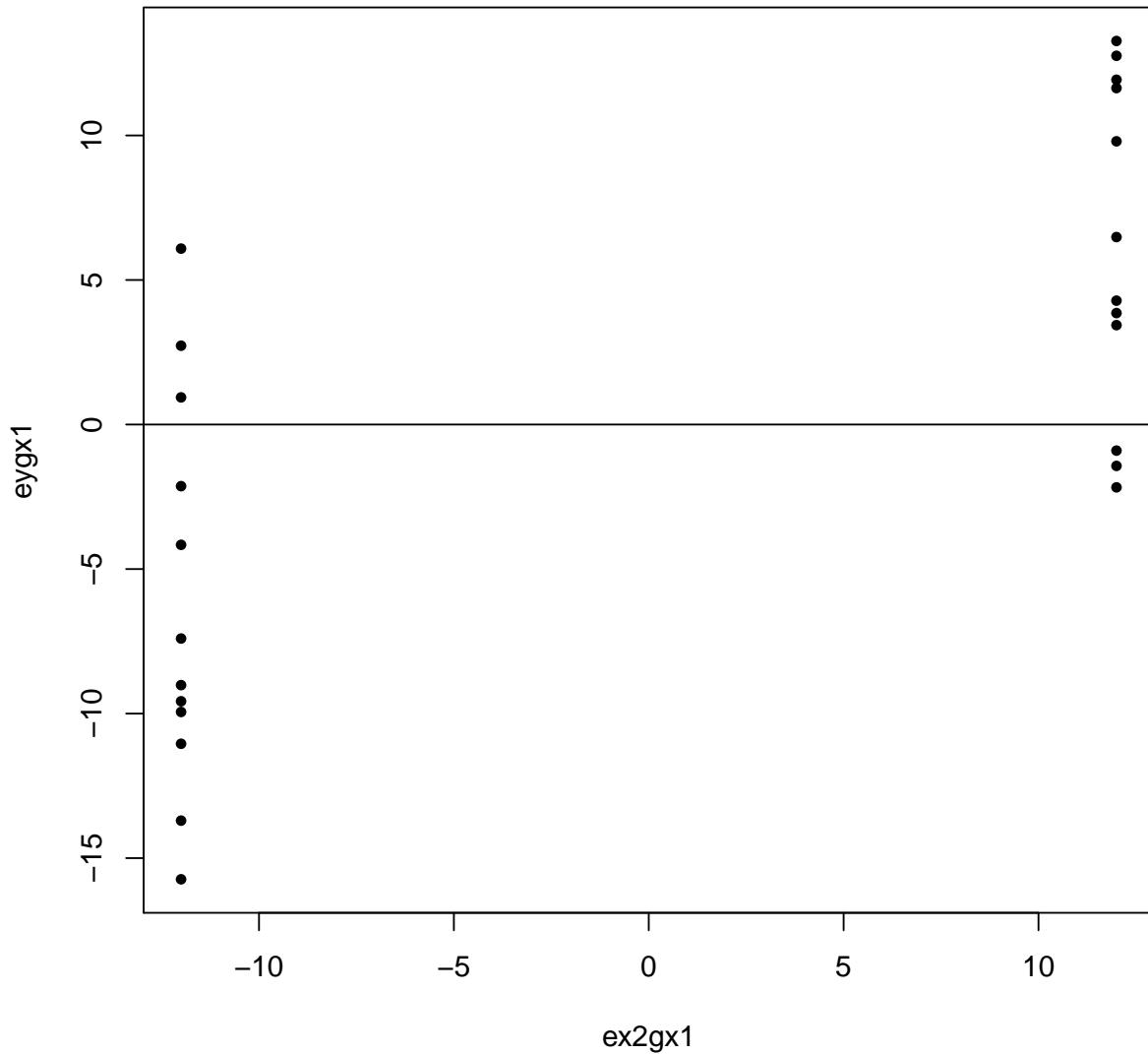
The following **omitted variable plot** suggests residual association with Days, as our EDA suggested, of course ([KNNL05, p. 112-3]).

```
> ### Illustration of omitted (not added) variable plot to diagnose an omitted  
> ### predictor:  
>  
> plot(residuals(flower1.lm) ~ Days, data=flower.df, pch=20)  
> abline(h=0)
```



Next is an **added variable plot** ([KNNL05, Sec. 10.1]), similar in spirit to the omitted variable plot. I prefer added variable plots.

```
> ex2gx1<- residuals(lm(Days ~ Intensity, data=flower.df))
> eygx1<- residuals(flower1.lm)
> plot(eygx1 ~ ex2gx1, pch=20)
> abline(h=0)
```



## 7.6 Remodeling

Of course, as our previous EDA suggested, the **omitted variable plot** and the **added variable plot** suggest that the number of flowers on a plant increases for those plants whose light regimen begins earlier, i.e., more days prior to PFI; i.e., Days is a so-called **omitted variable** or a variable to be added.

```
> ### Add Days (as a quantitative variable) to the model (Our first
> ### MLR!):
>
> flower2.lm<- lm(Flowers ~ Intensity + Days, data=flower.df)
> summary(flower2.lm)

Call:
lm(formula = Flowers ~ Intensity + Days, data = flower.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.652 -4.139 -1.558  5.632 12.165 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71.305833   3.273772 21.781 6.77e-16 ***
Intensity   -0.040471   0.005132 -7.886 1.04e-07 ***
Days        0.506597   0.109565  4.624 0.000146 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.441 on 21 degrees of freedom
Multiple R-squared:  0.7992, Adjusted R-squared:  0.78
F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08
```

## 7.7 Observed vs. Fitted Plot

The next plot is a common graphical display to informally assess the model fit.

```
> plot(flowers.df$Flowers ~ fitted(flower2.lm))

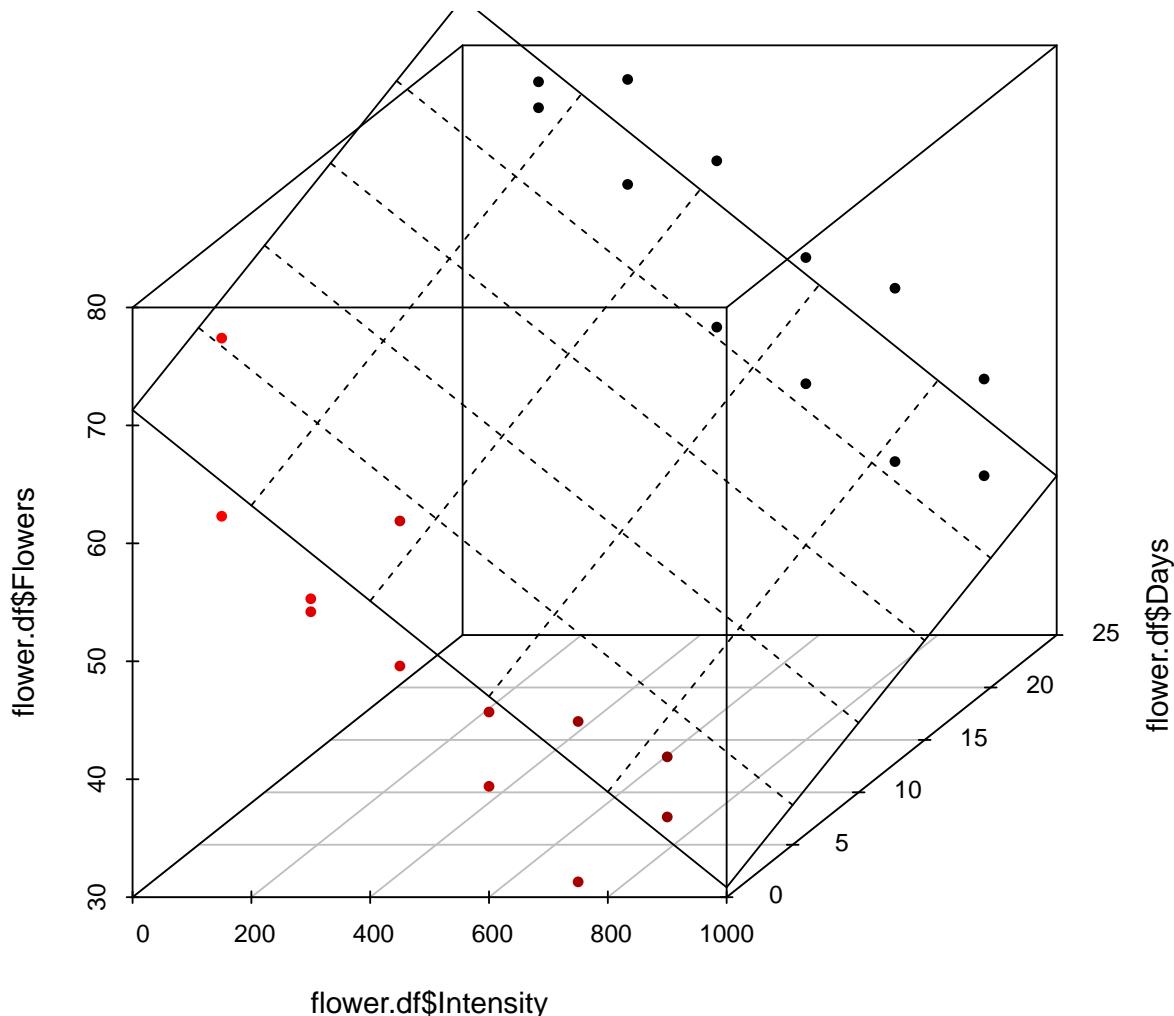
Error in eval(predvars, data, env): object 'flowers.df' not found

> abline(c(0,1))

Error in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...): plot.new has
not been called yet
```

The next plot shows the fitted plane and the observations, but is not typically called an “observed vs fitted” plot, and is less typical than the above plot.

```
> library(scatterplot3d)
> ### Another way to visually diagnose your model (good?).
> ### (See Display 9.5 in Ramsey and Schefer)
> flower.3d<- scatterplot3d(x=flower.df$Intensity, y=flower.df$Days,
+                             z=flower.df$Flowers, type=c("p"),
+                             pch=20, highlight.3d=TRUE)
> flower.3d$plane3d(flower2.lm, lty.box = "solid")
```



```
> detach(package:scatterplot3d)
```

## 7.8 Residual Plots

Generally speaking, residuals are useful for diagnosing a fitted model's departures from various model assumptions. We introduced "ordinary" residuals in

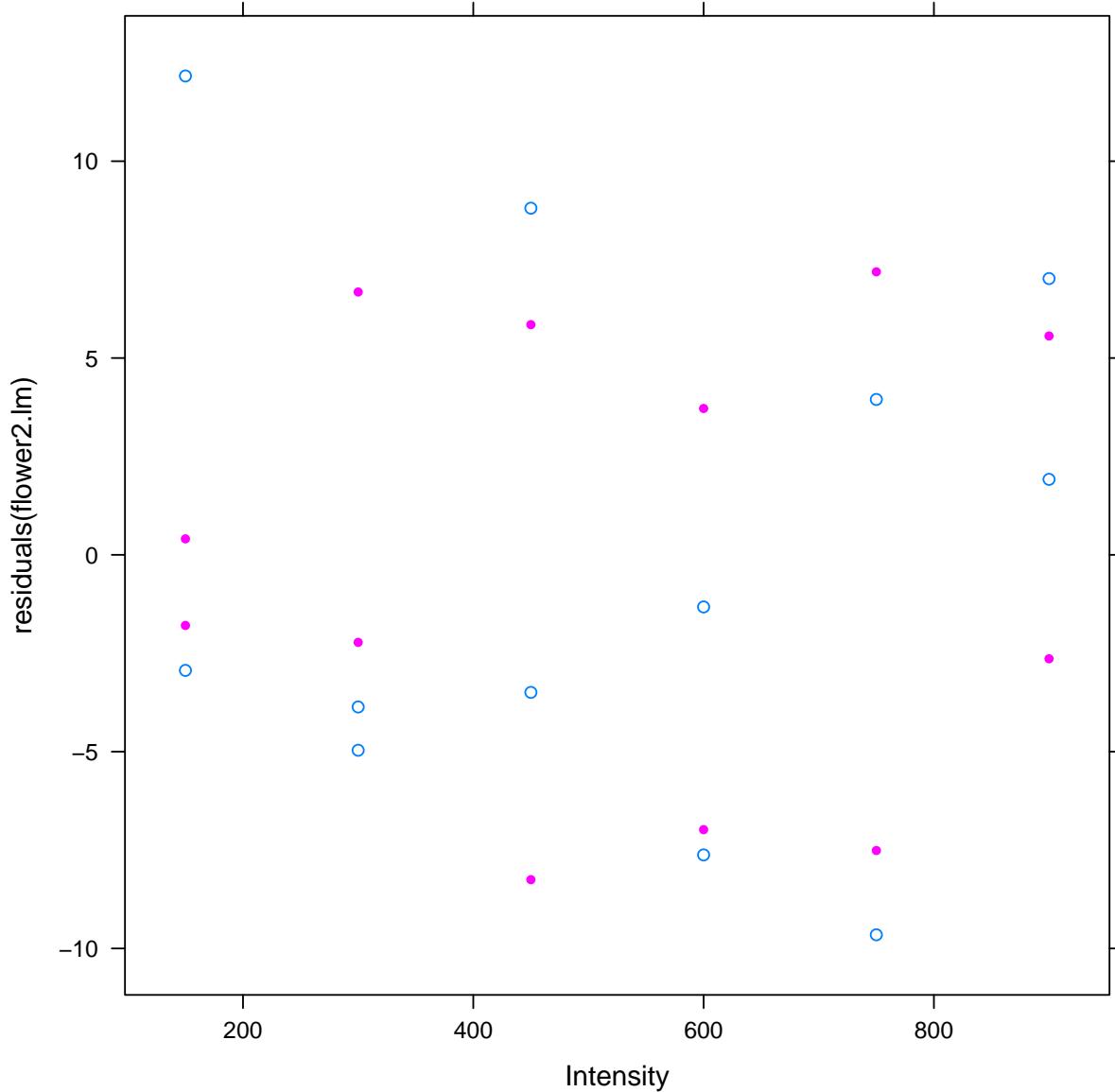
Definition 6.2. For now, we plot ordinary residuals vs. covariates (Intensity and Days).

Largely speaking, for the plot that we look at now, we should look at such plot in much the same way as we would a plot of response vs. covariates. In particular, we are looking to (re)model any (remaining) systematic relationship between the residuals and covariates, as we would with the original response and covariates. If there is such a systematic relationship or trend in the residuals, then this suggests that we have gotten wrong the division of variability in our response between the mean (systematic variability) and error (random variability), and we need to include some additional covariates or further function of the covariates in our mean model to shift variability from the residuals (“observed error”) into the (estimated) mean. In other words, our **diagnosis** of the residuals indicate that our model exhibits a **lack of fit** to the data, and we look to include other covariates or transformations of covariates as a **remedial measure**.

We might also look to these residual plots for other departures, such as non-constant variance, non-normality or outlying observations, but, as a rule of thumb, try to get diagnose the model correct first, then diagnose/remediate other departures, then revisit your mean model to see if anything has changed.

We will have much more to say about using residuals for diagnostics and remedial measures; see [Wak13, §5.11.2 and §5.11.3]. In particular, once we feel somewhat comfortable with a regression function model, we may use residuals to help guide (co)variance modeling beyond constant variance,  $\text{Var}(\epsilon_i) = \sigma^2$  (if we have time!).

```
> ### One of several ways to view residuals:
> library(lattice)
> xyplot(residuals(flower2.lm) ~ Intensity, groups=Days,
+         data=flower.df, pch=c(1,20))
```



```
> ##legend(x=500,y=12,legend=c("0 days","24 days"),pch=c(1,20))
> ##abline(h=0)
```

- What is your diagnosis, Dr.?

## 7.9 Overall F Test: Special Case of $\mathbf{C}\boldsymbol{\beta}$

### *Additional Reading:*

[Wak13, p. 213]

[RS13, pg 286, Subsection “Special Case:...” and Sec. 10.3.3]

[KNNL05, p. 226] \_\_\_\_\_  $\mathcal{R}$

- Here, we look at a special case of the special case (special special case?) (§6.7.5) of inferring  $\mathbf{C}\boldsymbol{\beta}$ , known as the **overall F-test**, which essentially compares a current model (alternative hypothesis or full model) to the constant mean model (null hypothesis or reduced model with all but regression function parameters except the intercept being zero).
- It is a test for any “overall” **linear** relationship of the response with the covariates, collectively. In other words, what happens if we throw out all of our covariates? We may have mentioned this test showing up in the default output of the `lm` function in the brief analysis of the Big Bang data of section 6.8, which was just a test for the slope, in that SLR.
- The overall F-test is **not often the object of direct interest**, and, contrary to what you may hear, a “non-significant” overall F-test does not mean that we cannot conduct further, more detailed and interesting inferences.

- Incidentally, an F-test may be called a **partial** F-test when testing whether to throw out one covariate (or a few; less than all) ([Wak13, p. 212]).

- We often see, traditionally, at least, the overall (and other) F-test summarized very methodically in an “**ANOVA table**”.
- ANOVA stands for “ANalysis Of VAriance” and simply means that we break apart (analyze) the total sum of squared deviations,

$$TSS = \sum_i (y_i - \bar{y})^2$$

([Wak13, p. 213]) into pieces that we attribute to one or more covariates, then use these pieces in the construction of F-tests for the various covariates. (Or, it may be called CTSS ([Wak13, §4.8.2]) or SSTO or similar.)

- Incidentally, [Wak13, p. 211] attempts to decompose  $TSS$  into a residual sum of squares (RSS) and a fitted sum of squares (FSS (a difference between a full model RSS and a very reduced model RSS, by the way)), but something seems to have gone awry (more than just a **typo?**).
- For the overall F-test for regression, see, generically, [KNNL05, Sec. 6.5] and, with numbers for our flower data, [RS13, Display 10.11].
- The **method known as ANOVA** (often distinguished from regression, though its just a special case of linear regression model) also traditionally presents results in (more interesting) ANOVA tables.
- **We will not focus on ANOVA tables.** There are many books that have and still do. I tend to think of them as a hold-over from when computations were done by hand (in certain “balanced” data

situations), and results were neatly summarized in a table. I don't think we need them.

### Overall F-test Hypotheses

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

$$H_a : \text{not all } \beta_j, j = 1, \dots, p-1 \text{ are zero}$$

Notice  $\beta_0$  is not part of this test.

- How do we write these hypotheses in the matrix terms of the General Linear Hypothesis of §6.7.2? (Assume the currently fitted model in our running example.)
- From perspective of the “Full vs. Reduced” model approach of §6.7.4, what is the full model? The reduced model?
- Do you see the connection between the hypotheses and the F/R models?

### 7.9.1 anova Function for F v R Approach

- We could use the “by hand”  $\mathbf{C}\boldsymbol{\beta}$  matrix approach as in the code of the Big Bang example of §6.8; the code would be nearly the same as in that section. (Perhaps for a homework.)
- Instead, we get the overall F test from the **default printout** (as we may have mentioned briefly in the Big Bang code) R

- And, we will use the `anova` function to implement the “F v R” approach (§6.7.4), which we have not implemented before.
- I feel like we should perform the “F v R” approach “by hand,” too (get  $\text{RSS}(\text{Full})$ ,  $\text{RSS}(\text{Reduced})$ , etc.); perhaps for homework.

```
> ## Default printout:
> summary(flower2.lm)

Call:
lm(formula = Flowers ~ Intensity + Days, data = flower.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.652 -4.139 -1.558  5.632 12.165 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71.305833   3.273772 21.781 6.77e-16 ***
Intensity   -0.040471   0.005132 -7.886 1.04e-07 ***
Days        0.506597   0.109565  4.624 0.000146 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.441 on 21 degrees of freedom
Multiple R-squared:  0.7992, Adjusted R-squared:  0.78 
F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08

> alpha<- 0.05; n<-24; p<-3
> qf(1-alpha,p-1,n-p)

[1] 3.4668

> ## F v R approach:
> flower2R.lm<- update(flower2.lm, . ~ . - Intensity - Days)
> summary(flower2R.lm)
```

```
Call:  
lm(formula = Flowers ~ 1, data = flower.df)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-24.837 -10.713 -1.387  8.312 21.863  
  
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 56.137     2.803   20.02 4.7e-16 ***  
---  
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 13.73 on 23 degrees of freedom  
  
> anova(flower2R.lm, flower2.lm)  
  
Analysis of Variance Table  
  
Model 1: Flowers ~ 1  
Model 2: Flowers ~ Intensity + Days  
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)  
1      23 4337.9  
2      21  871.2  2    3466.7 41.78 4.786e-08 ***  
---  
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- To discuss: test statistic? p-value? critical value (Type I error probability?) of the test? conclusion?

### 7.9.2 The `glh.test` R Function

- Here, we introduce a function for inferring about general linear combinations of the form  $\mathbf{C}\boldsymbol{\beta}$ , which we've computed in various ways for particular examples.
- Of course, we have our “by hand”  $\mathbf{C}\boldsymbol{\beta}$  code for this, but, in practice, we typically want more convenience.
- We might think to use the `anova` function to implement the “F v R” approach, as we illustrated, above. After all, that approach seems easy. But, sometimes we may have a bit of difficulty figuring out what the reduced model is that corresponds to the null hypothesis (not in any of the previous examples). The function will come in handy, later, for ANOVA inferences.
- The tool is the `glh.test` function in the `gmodels` package, which you may have to download and install. We illustrate its use briefly by repeating the overall F-test, which we did above using the default printout and the `anova` function. Again, `glh.test` is used for much more than just the overall F-test.

```
> library(gmodels)
> Cmat<- matrix(c(0,1,0,
+                  0,0,1), ncol=3, byrow=TRUE)
> colnames(Cmat)<- c("(Intercept)","Intensity","Days")
> b0<- c(0,0)
> glh.test(flower2.lm, cm=Cmat, d=b0)

Test of General Linear Hypothesis
Call:
glh.test(reg = flower2.lm, cm = Cmat, d = b0)
F = 41.7801, df1 = 2, df2 = 21, p-value = 4.786e-08

> detach(package:gmodels)
```

## 7.10 $R^2$ & Adjusted $R^2$

**Additional Reading:**

[Wak13, §4.8.2]

[RS13, §8.6.1, 10.4.1]

[KNNL05, pp. 226-7] \_\_\_\_\_  $\mathcal{R}$

- Here, we use the “RSS” notation of our “F v R” approach (§6.7.4). We give full model’s RSS a special label,

$$SSE = RSS(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{x}^T \hat{\boldsymbol{\beta}})^T (\mathbf{Y} - \mathbf{x}^T \hat{\boldsymbol{\beta}}).$$

(Note this is the numerator of  $MSE$  (Definition 6.3).)

- And, we consider a very reduced model RSS, with a special label,

$$\begin{aligned} TSS &= RSS(\beta_0) = \sum_i (y_i - \hat{\beta}_0)^2 \\ &= \sum_i (y_i - \bar{y})^2 \\ &= (\mathbf{y} - \bar{y}\mathbf{1})^T (\mathbf{y} - \bar{y}\mathbf{1}) \end{aligned}$$

(the estimate of  $\beta_0$  in each model is generally different, of course!).

- And, we denote the difference in these models’ RSS values, called the **regression sum-of-squares**:

$$SSR = TSS - SSE$$

([Wak13, pp. 212-3] might denote this difference with his “fitted sum-of-squares” notation, something like  $FSS(\boldsymbol{\beta}_{-0} | \beta_0) = RSS(\beta_0) - RSS(\boldsymbol{\beta})$ ).

**Definition 7.1** (Coefficient of Determination ( $R^2$ )).

$$R^2 = \frac{SSR}{TSS} = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS}$$

([Wak13, §4.8.2]). Interpretation : The proportion of total variability ( $TSS$ ) in the response observations associated with the linear regression on the covariates  $x_1, \dots, x_k$ .

- For the flower example, we have

$$R^2 = ???$$

- Thus, ???% of the variability in the number of flowers on a plant is explained by the *linear* association with light intensity and the number of days prior to PFI that the light regimen was begun.
- $R^2$  is often presented as one of several **model selection** criteria: bigger = better. But, better and more popular selection/building criteria exist.
- Incidentally, the **coefficient of multiple correlation** is

$$R = \sqrt{R^2} \quad (\text{positive root!})$$

$R = |r|$  where  $r$  is the **correlation coefficient** between  $y_i$  and a single observed covariate,  $x_i$ , as in simple linear regression (SLR).

**Definition 7.2** (Adjusted  $R^2$ ).

$$\begin{aligned} R_a^2 &= \frac{\frac{TSS}{n-1} - \frac{SSE}{n-p}}{\frac{TSS}{n-1}} \\ &= 1 - \frac{\frac{SSE}{n-p}}{\frac{TSS}{n-1}} \\ &= 1 - \left( \frac{n-1}{n-p} \frac{SSE}{TSS} \right) \\ &= 1 - (1 - R^2) \left( \frac{n-1}{n-p} \right) \end{aligned}$$

([Wak13, §4.8.2]). Thus, we see the adjusted  $R^2$  “penalizes” a larger number of parameters,  $p$ , in the regression model, thus somehow encouraging **parsimony** in model (variable) selection compared to ordinary  $R^2$ , which never decreases as we add more covariates, even if the covariates are not statistically significant.

For the flower example, we have

$$R_a^2 = ???$$

## 7.11 Using an Interaction Term

**Additional Reading:**

[RS13, Sec. 9.3.4]

[KNNL05, pp. 220 & Sec. 8.2] \_\_\_\_\_  $\mathcal{R}$

Illustrating with a model for flowers, consider

$$Y_i = \beta_0 + \beta_1 \overbrace{x_{i1}}^{\text{Intensity}} + \beta_2 \overbrace{x_{i2}}^{\text{Days}} + \beta_3 \overbrace{x_{i1}x_{i2}}^{\text{interaction}} + \epsilon_i,$$

with the usual assumptions. Notice how the effect on the mean response to (effect of) one interaction term variable depends on the other interaction term variable (see §5.5.3). For example, fix  $x_1$  and  $x_2$  at some values, then investigate the effect on the mean with a one unit increase in, say,  $x_1$ :

$$\begin{aligned} & \beta_0 + \beta_1(x_1 + 1) + \beta_2x_2 + \beta_3(x_1 + 1)x_2 \\ & - (\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2) \\ & \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \\ & = 0 + \beta_1 + 0 + \beta_3x_2 \end{aligned}$$

In other words, the change in mean response with a 1 unit increase in  $x_1$  depends on the value of  $x_2$ , i.e., the rate of change (partial derivative) of the mean number of flowers with light intensity depends on when the light regimen begins. We fit this interaction model in the next Chunk.

NOTE: Some statisticians argue that if you have a second order interaction  $x_1x_2$ , then you should put into the model the corresponding squared terms  $x_1^2$  and  $x_2^2$ , then test whether all three 2nd order polynomial terms (those involving  $x_1^2$ ,  $x_2^2$  and  $x_1x_2$ ) are simultaneously significant (using the full vs. reduced model F-test of course). This stems from the **marginality principle** or what [Wak13, p. 205] calls the **hierarchy principle**. We merely issue a warning, and mention this principle is typically not of concern for predictive modeling ([Wak13, Part IV]).

```
> ### Does the rate of change (slope) of flowering (Flowers) with light
> ### intensity (Intensity) depend on the onset time (Days) of light
> ### treatment?:
>
> ### Example of an interaction term:
> flower3.lm<- lm(Flowers ~ Intensity + Days + Intensity:Days,
```

```

+           data=flower.df)
> summary(flower3.lm)

Call:
lm(formula = Flowers ~ Intensity + Days + Intensity:Days, data = flower.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.516 -4.276 -1.422  5.473 11.938 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71.6233333  4.3433046 16.491 4.14e-13 ***
Intensity   -0.0410762  0.0074351 -5.525 2.08e-05 ***
Days        0.4801389  0.2559317  1.876  0.0753 .  
Intensity:Days 0.0000504  0.0004381  0.115  0.9096  
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 6.598 on 20 degrees of freedom
Multiple R-squared:  0.7993, Adjusted R-squared:  0.7692 
F-statistic: 26.55 on 3 and 20 DF,  p-value: 3.549e-07

```

Plugging in estimates from the previous Chunk, we have the rate of change

$$-0.0410762 + 0.00005404x_2,$$

which is

$$-0.0410762 \quad (x_2 = 0 \text{ days prior to PFI})$$

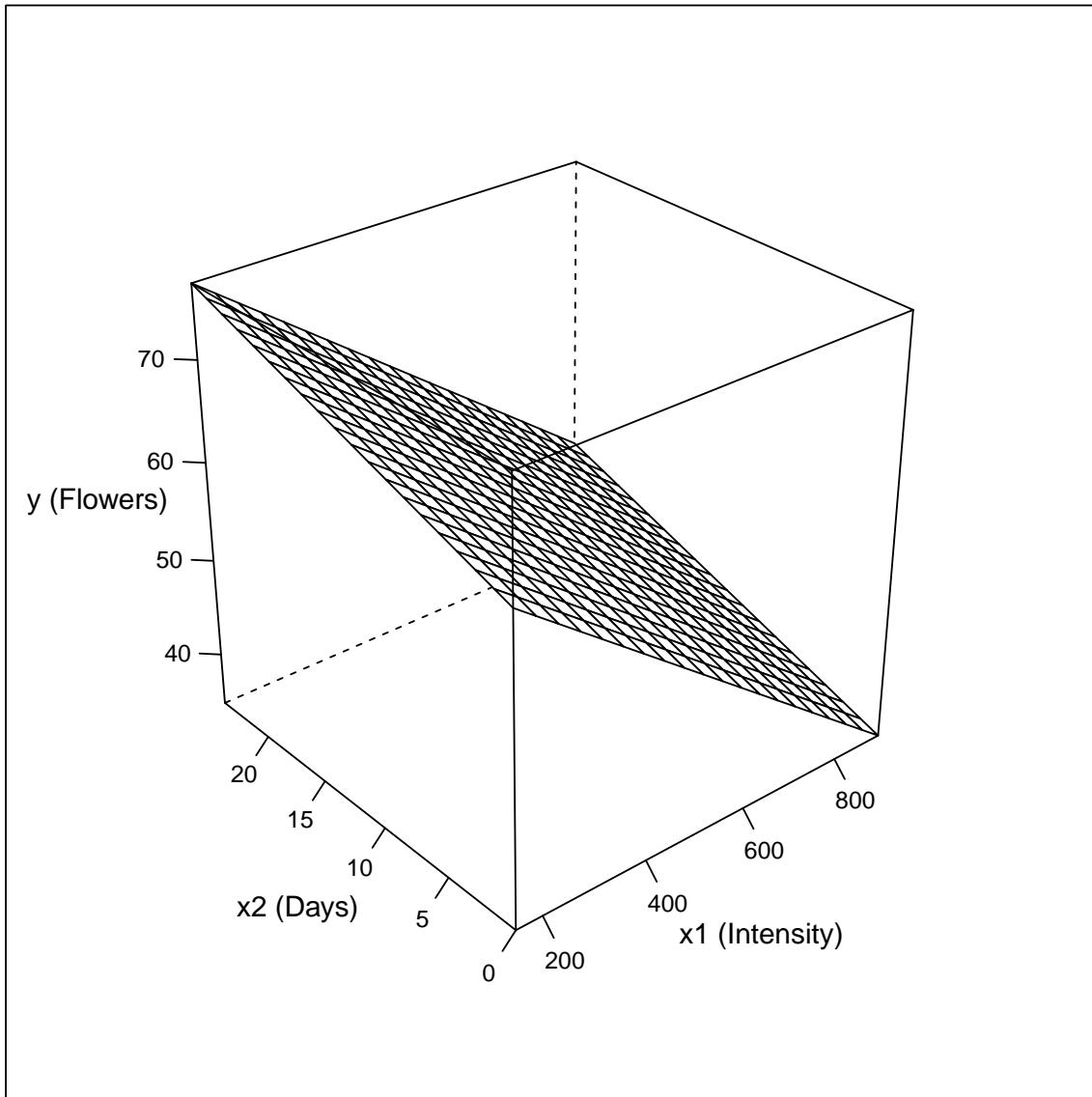
and,

$$-0.039779 \quad (x_2 = 24 \text{ days prior to PFI}).$$

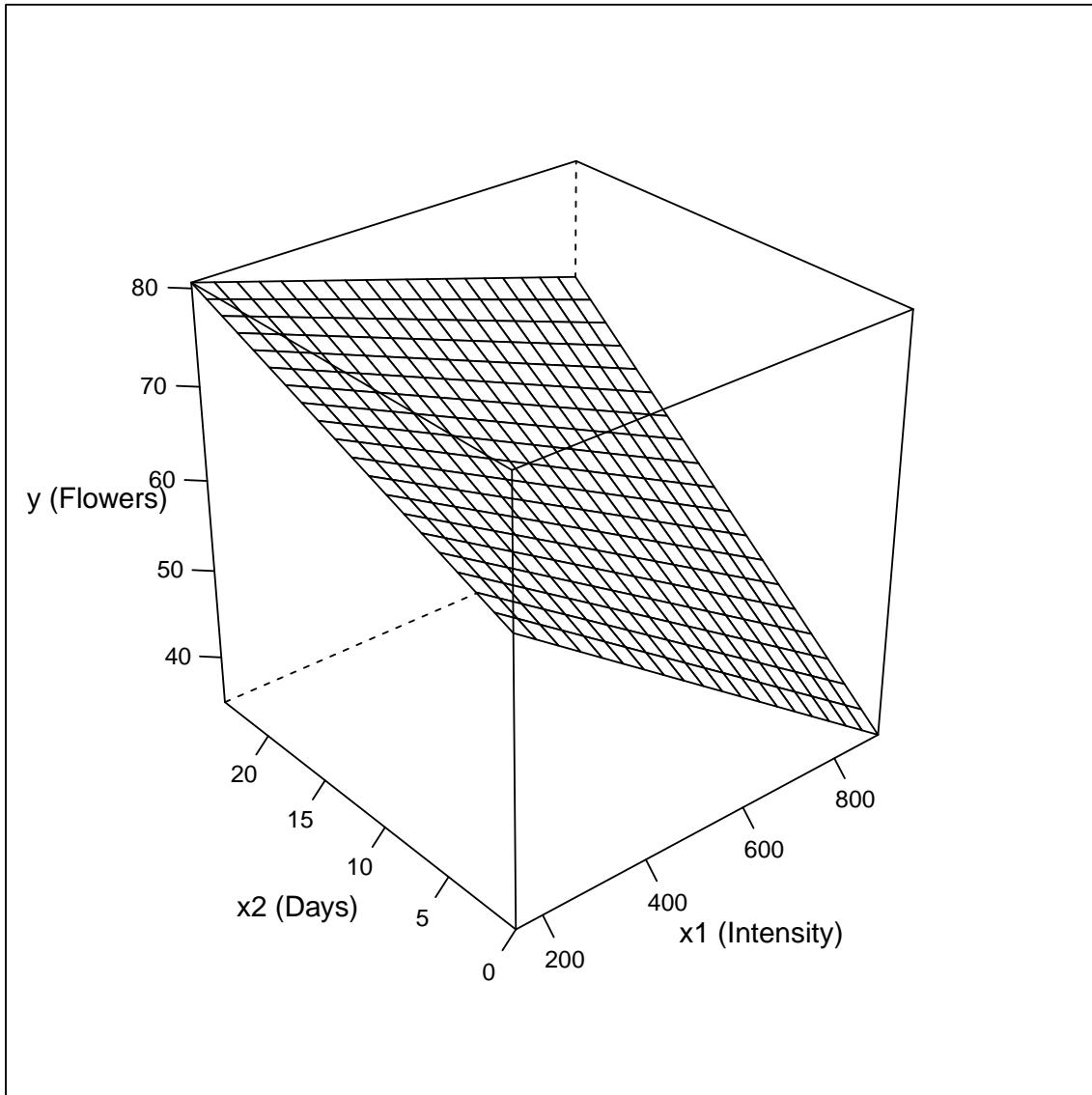
Thus, it appears as if we get a lower rate of decrease in mean number of flowers when starting the light regimen 24 days prior to PFI. Why might we be cautious about inferring for, say,  $x_2 = 12$  days?

The following two Chunks illustrate respectively the fitted and a hypothetical regression or response surface with interactions. More discussion in class.

```
> ### Estimated flower interaction surface
> minx1<- min(flower.df$Intensity); maxx1<- max(flower.df$Intensity)
> x1grid<- seq(minx1, maxx1,length=20)
> minx2<- min(flower.df$Days); maxx2<- max(flower.df$Days)
> x2grid<- seq(minx2, maxx2,length=20)
> x1x2grid<- expand.grid(x1=x1grid, x2=x2grid)
>
> betas<- coef(flower3.lm)
>
> y<- betas[1] + betas[2]*x1x2grid$x1 + betas[3]*x1x2grid$x2 +
+     betas[4]*x1x2grid$x1*x1x2grid$x2
> library(lattice)
> wireframe(y ~ x1x2grid$x1 + x1x2grid$x2, xlab="x1 (Intensity)",
+             ylab="x2 (Days)", zlab="y (Flowers)",
+             scales=list(arrows=FALSE),
+             main="Flowers: Fitted Surface")
```

**Flowers: Fitted Surface**

```
> ### Hypothetical flower interaction surface to exaggerate
> ### dependence of rate of change of flowers with Intensity on Days.
> ### (This is not a plot of a fitted model. Just for illustration.)
> y<- betas[1] + betas[2]*x1x2grid$x1 + betas[3]*x1x2grid$x2 +
+   0.001*x1x2grid$x1*x1x2grid$x2
> wireframe(y ~ x1x2grid$x1 + x1x2grid$x2, xlab="x1 (Intensity)",
+            ylab="x2 (Days)", zlab="y (Flowers)",
+            scales=list(arrows=FALSE),
+            main="Flowers: Hypothetical Mean Surface")
```

**Flowers: Hypothetical Mean Surface**

```
> ##detach(package:lattice)
```

## 7.12 $t$ -based inference for $\beta_j$

We again illustrate  $t$ -based inferences for individual parameters,  $\beta_j$ , based on previous results (§6.7.7 & 6.7.8). We should be able to write a nice summary of a test or CI for each coefficient, component-wise or with matrices. More in class.

### 7.12.1 Default lm Printout and summary

The default printout summarizes 2-sided t-tests for null values of zero.

```
> flower2.lm

Call:
lm(formula = Flowers ~ Intensity + Days, data = flower.df)

Coefficients:
(Intercept)      Intensity          Days
           71.30583     -0.04047       0.50660

> summary(flower2.lm)

Call:
lm(formula = Flowers ~ Intensity + Days, data = flower.df)

Residuals:
    Min     1Q Median     3Q    Max 
-9.652 -4.139 -1.558  5.632 12.165 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71.305833   3.273772  21.781 6.77e-16 ***
Intensity   -0.040471   0.005132  -7.886 1.04e-07 ***
Days        0.506597   0.109565   4.624 0.000146 ***
---

```

```

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.441 on 21 degrees of freedom
Multiple R-squared:  0.7992, Adjusted R-squared:  0.78
F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08

```

### 7.12.2 By Hand Test and Intervals

Here, we revert to “hand” computations to draw connections to the higher level R that you would typically use in practice. We

```

> ## Estimated variance matrix of regression parameters:
> (flower2.lm.vcov<-vcov(flower2.lm))

      (Intercept)       Intensity        Days
(Intercept) 10.71758322 -1.382914e-02 -1.440535e-01
Intensity    -0.01382914  2.634122e-05  1.699154e-20
Days         -0.14405354  1.699154e-20  1.200446e-02

> betahat<- coef(flower2.lm)
> beta0<- rep(0,3) ## e.g.
> sehat<- sqrt(diag(flower2.lm.vcov))
> tstar<- (betahat-beta0)/sehat
>
> n<-24; p<-3
> pval<- 2*pt(abs(tstar),n-p, lower.tail=FALSE)
>
> cbind(estimate=betahat, se=sehat, tstar=tstar, pval2side=pval)

      estimate          se      tstar     pval2side
(Intercept) 71.30583333 3.27377202 21.780940 6.767274e-16
Intensity   -0.04047143 0.00513237 -7.885525 1.036787e-07
Days        0.50659722 0.10956487  4.623719 1.463776e-04

> alpha<- 0.05
> tcrit<- qt(1-alpha/2, n-p)
> cbind(lb=beta0-tcrit*sehat, ub=beta0+tcrit*sehat)

      lb          ub
(Intercept) 64.49765172 78.11401495
Intensity   -0.05114478 -0.02979808
Days        0.27874459 0.73444985

```

### 7.12.3 The confint Function

The default confidence level is 0.95.

```
> confint(flower2.lm, level=0.95)

            2.5 %      97.5 %
(Intercept) 64.49765172 78.11401495
Intensity    -0.05114478 -0.02979808
Days         0.27874459  0.73444985
```

### 7.12.4 The estimable Function

The `estimable` function in the `gmodels` package is often useful for inferring about  $\beta_j$ . `glh.test` is in the same package. Again, you may have to download and install the `gmodels` package. **NOTE:** By default, the `estimable` function gives individual tests/CIs, not a simultaneous or joint F-test like `glh.test`. See the `joint.test` argument of `estimable`. More in class.

```
> library(gmodels)
> estimable(flower2.lm, cm=diag(3), beta0=c(0,0,0), conf.int=0.95)

      beta0   Estimate Std. Error   t value DF
(1 0 0)     0 71.30583333 3.27377202 21.780940 21
(0 1 0)     0 -0.04047143 0.00513237 -7.885525 21
(0 0 1)     0  0.50659722 0.10956487  4.623719 21
      Pr(>|t|) Lower.CI   Upper.CI
(1 0 0) 6.661338e-16 64.49765172 78.11401495
(0 1 0) 1.036787e-07 -0.05114478 -0.02979808
(0 0 1) 1.463776e-04  0.27874459  0.73444985

> ## Illustrate non-zero null values (not that these are interesting tests):
> estimable(flower2.lm, cm=diag(3), beta0=c(0,1,1), conf.int=0.95)

      beta0   Estimate Std. Error   t value DF
(1 0 0)     0 71.30583333 3.27377202 21.780940 21
(0 1 0)     1 -0.04047143 0.00513237 -202.727297 21
(0 0 1)     1  0.50659722 0.10956487  -4.503293 21
      Pr(>|t|) Lower.CI   Upper.CI
(1 0 0) 6.661338e-16 64.4976517 78.1140149
(0 1 0) 0.000000e+00 -1.0511448 -1.0297981
(0 0 1) 1.950801e-04 -0.7212554 -0.2655502
```

```
> detach(package:gmodels)
```

## 7.13 Qualitative Covariates

### *Additional Reading:*

[Wak13, §5.5.2]

[RS13, Sec. 9.3 and bat e.g. throughout Chap. 10]

[KNNL05, pp. 218-9 & Sec. 8.3 & 8.4] \_\_\_\_\_  $\mathcal{R}$

We should already know what a qualitative variable is.

- Color: red, green, blue (nominal)
- Gender: male, female (nominal)
- Height class: short, average, tall (ordinal)
- Flowers e.g. Time: 0 days since PFI, 24 days since PFI (ordinal).  
IMPORTANT: we treat “0 days” and “24 days” as labels, with no quantitative interpretation. We will even ignore the ordinal nature of these variables “values”. (Recall, we created the quantitative “Days” variable.)
- a.k.a, **factor** or **categorical variable**

### 7.13.1 Cell Reference Coding

How do we do regression with categorical valued  $x$  variables, i.e., with factors? Numerical codes. There are **many ways to code** categorical

variables. Each will give a different parameterization and interpretation of the regression model! Here, we cover **cell reference coding** (or reference cell) or **treatment coding** or, as [Wak13, §5.5.2] calls it, **corner-point coding**. This is the default in R (which “cell” or which “corner”?)

If we have a categorical variable (factor) with

$k$  categories (possible variable values, i.e., levels),

then we create

$(k - 1)$  new quantitative predictor variables

in the following manner.

- Say,  $X$  is a categorical variable with  $k = 3$  levels: “level 1”, “level 2”, and “level 3”.
- Create  $3-1 = 2$  new predictor variables

$$X_1 = \begin{cases} 1 & \text{if } X = \text{level 2} \\ 0 & \text{if } X \text{ value is not level 2} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if } X = \text{level 3} \\ 0 & \text{if } X \text{ value is not level 3} \end{cases}$$

- Then do regression using the new predictor variables!

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \text{other terms} + \epsilon_i$$

The “design matrix”, i.e.,  $\mathbf{X}$  will have  $(k - 1)$  columns for these new regressors (predictors) (among other columns for other regressors), e.g.,

| $X_1$ | $X_2$ | original $X$ |
|-------|-------|--------------|
| 0     | 0     | level 1      |
| 0     | 0     | level 1      |
| 0     | 0     | level 1      |
| 1     | 0     | level 2      |
| 1     | 0     | level 2      |
| 1     | 0     | level 2      |
| 0     | 1     | level 3      |
| 0     | 1     | level 3      |
| 0     | 1     | level 3      |

- NOTE: Again, this is not the only way to code categorical variables.

**Interpretation of parameters** for the (cell reference) coded variables.

- $\beta_1$  is the shift (up/down) in the regression function (surface, hyperplane) from reference  $X = \text{level 1}$  to  $X = \text{level 2}$
- $\beta_2$  is the shift (up/down) in the regression function (surface, hyperplane) from reference  $X = \text{level 1}$  to  $X = \text{level 3}$
- $\beta_0$  corresponds to the “reference level” (or “corner-point”) (level 1 here). Level 1 is the level to which all other levels of original  $X$  are compared. A convenient reference level situation in practice arises when there is a “control” level as in an experiment with a “control treatment” level with several other “treatment” levels. See “Redefining the Reference Level” [RS13, pg 280].
- IMPORTANT: Again, there are several ways to code categorical variables, and parameter INTERPRETATION DEPENDS ON CODING! See [KNNL05, 8.4]

Continuing with our flower example, now, we use Time as a categorical covariate i.e., a factor, instead of using Days as a quantitative covariate (we may see this in a homework too). For Time, we have only  $k = 2$  levels, so we need only  $k - 1 = 1$  coded covariate. Time=“0 days” is coded with a value of 0 and Time=“2 days” is coded with a value of 1. This simply creates a column in the  $\mathbf{X}$  matrix corresponding to a covariate that assumes values 0 and 1.

Incidentally, [RS13] cover this example (without numbers!) in their §9.3.3 and illustrate it in the second plot of their Display 9.8.

In general, we need to tell R what variables are factors, then how to

code each of these factors. Or, at least, we should know how R will treat covariates! We will spend much more time on coding factors when we get to ANOVA. Back in §7.2, we already prepared Time as a categorical variable, which R continues to recognize in the next chunk. In particular, R sees Time sees the level of Time as **ordered**.

We do not use this extra information, but, for our illustration, we still want to ensure that we get **cell reference (treatment, corner-point)** coding. We could (1) change the class of the factor to “factor” (remove ordered class) to use the default coding (if the default has not been changed!); (2) add an attribute to the factor to indicate what type of coding (“contrasts”) we want for that factor; or (3) change the global default for coding factors. NOTE: if a factor has a coding or “contrasts” attribute, it will override the global setting.

In our illustration, in a very real sense, the result is no different than in our previous analysis when using Intensity and Days (quantitative) as covariates (§7.6), but the parameter  $\beta_2$ , for Time, is now interpreted somewhat differently than it was for Days.

```
> is.factor(flower.df$Time)
[1] TRUE

> class(flower.df$Time)
[1] "ordered" "factor"

> levels(flower.df$Time)
[1] "0 days"   "24 days"

>getOption("contrasts")
[1] "contr.treatment" "contr.treatment"

>options(contrasts=c("contr.treatment", "contr.treatment"))
>flower4.lm<- lm(Flowers ~ Intensity + Time,
+                   data=flower.df)
>summary(flower4.lm)
```

```

Call:
lm(formula = Flowers ~ Intensity + Time, data = flower.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.652 -4.139 -1.558  5.632 12.165 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71.305833   3.273772  21.781 6.77e-16 ***
Intensity   -0.040471   0.005132  -7.886 1.04e-07 ***
Time24 days 12.158333   2.629557   4.624 0.000146 ***  
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 6.441 on 21 degrees of freedom
Multiple R-squared:  0.7992, Adjusted R-squared:  0.78 
F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08

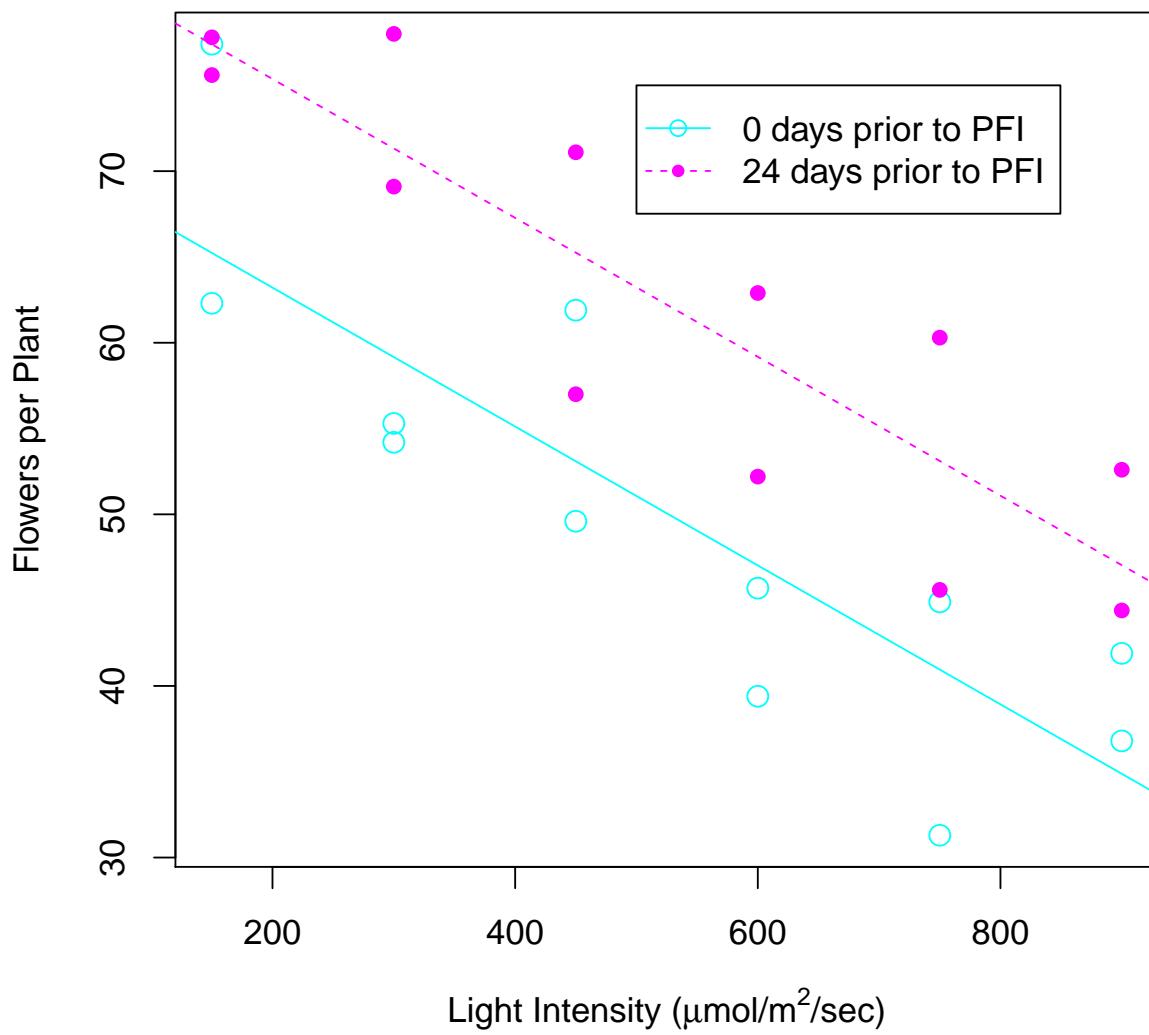
```

```

> attach(flower.df)
> par(cex=1.2)
> plot(Flowers[Time=="0 days"] ~ Intensity[Time=="0 days"],
+       xlab=expression(paste("Light Intensity (", mu, "mol/", m^2, "/sec)", 
+                             sep="")),
+       ylab="Flowers per Plant",
+       main="Meadowfoam Flower Experiment", pch=1, col="cyan", cex=1.3)
> points(Flowers[Time=="24 days"] ~ Intensity[Time=="24 days"], pch=20,
+         col="magenta", cex=1.3)
> legend(x=500,y=75,legend=c("0 days prior to PFI",
+                           "24 days prior to PFI"),
+         pch=c(1,20), col=c("cyan","magenta"), lty=c(1,2))
>
> ##### add fitted lines to plot:
> ab<- coef(flower4.lm)
> ab0days<- ab[c(1,2)] # int and slope for 0 days
> ab24days<- ab0days + c(ab[3],0) #int and slope for 24 days
> abline(ab0days, lty=1, col="cyan")
> abline(ab24days, lty=2, col="magenta")

```

## Meadowfoam Flower Experiment



```
> detach(flower.df)
```

## 7.14 Intervals for $E(Y|x)$ & $Y|x$ with predict

**Additional Reading:**

[RS13, 7.4.2, 7.4.3] for SLR

[RS13, 10.2.3, 10.2.4] for MLR

[KNNL05, 2.4, 2.5] for SLR

[KNNL05, 6.7] for MLR

$\mathcal{R}$

We use the model with Intensity and Time (factor), to illustrate Results 6.14 & 6.7.9. NOTE: Typically, in any given application, we usually perform one or the other, not both, depending on the goal. We skip the “hand” computations and use the `predict` function. See the Big Bang example in §6.8 for an illustration of the “hand” computations.

```
> ### Estimate  $E(Y/x)$  ( $C = x^t$  is a row vector now)
>
> ### E.g. Use flower4.lm to estimate mean number of flowers at
> ### light intensity 350 beginning 0 days prior to PFI
> ### i.e.,  $E(Y/x=(350, "0\ days"))$ :
>
> x<- data.frame(Intensity=350, Time=factor(1,levels=c(1,2),
+                                         labels=c("0\ days","24\ days")))
>
> predict(flower4.lm, newdata=x, se.fit=TRUE, interval="confidence",
+         level=0.95)

$fit
      fit      lwr      upr
1 57.14083 52.84655 61.43511

$se.fit
[1] 2.064942

$df
```

```
[1] 21

$residual.scale
[1] 6.441073

> ### NOTE: x ACTUALLY is c(1,350,0)^T, right!
>
> ### E.g. Use flower4.lm to estimate mean number of flowers at
> ### light intensity 350 beginning 24 days prior to PFI:
>
> x<- data.frame(Intensity=350, Time=factor(2,levels=c(1,2),
+                                         labels=c("0 days","24 days")))
>
> predict(flower4.lm, newdata=x, se.fit=TRUE, interval="confidence",
+         level=0.95)

$fit
      fit      lwr      upr
1 69.29917 65.00489 73.59345

$se.fit
[1] 2.064942

$df
[1] 21

$residual.scale
[1] 6.441073

> ### NOTE: x ACTUALLY is c(1,350,1)^T
>
> ### _Predict_ an unobserved number of flowers Y/x:
>
> ### E.g. Use flower4.lm to _predict_ Y/x=(350, "0 Days")
> ### (at light intensity 350 beginning 0 days prior to PFI):
>
> x<- data.frame(Intensity=350, Time=factor(1,levels=c(1,2),
+                                         labels=c("0 days","24 days")))
>
> predict(flower4.lm, newdata=x, se.fit=TRUE, interval="prediction",
+         level=0.95)

$fit
```

```

      fit      lwr      upr
1 57.14083 43.07437 71.2073

$se.fit
[1] 2.064942

$df
[1] 21

$residual.scale
[1] 6.441073

> ### NOTE: x ACTUALLY is c(1,350,0)^T
>
> ### E.g. Use flower4.lm to predict Y/x=(350, "24 Days")
> ### (at light intensity 350 beginning 24 days prior to PFI):
>
> x<- data.frame(Intensity=350, Time=factor(2,levels=c(1,2),
+                                         labels=c("0 days","24 days")))
>
> predict(flower4.lm, newdata=x, se.fit=TRUE, interval="prediction",
+         level=0.95)

$fit
      fit      lwr      upr
1 69.29917 55.2327 83.36563

$se.fit
[1] 2.064942

$df
[1] 21

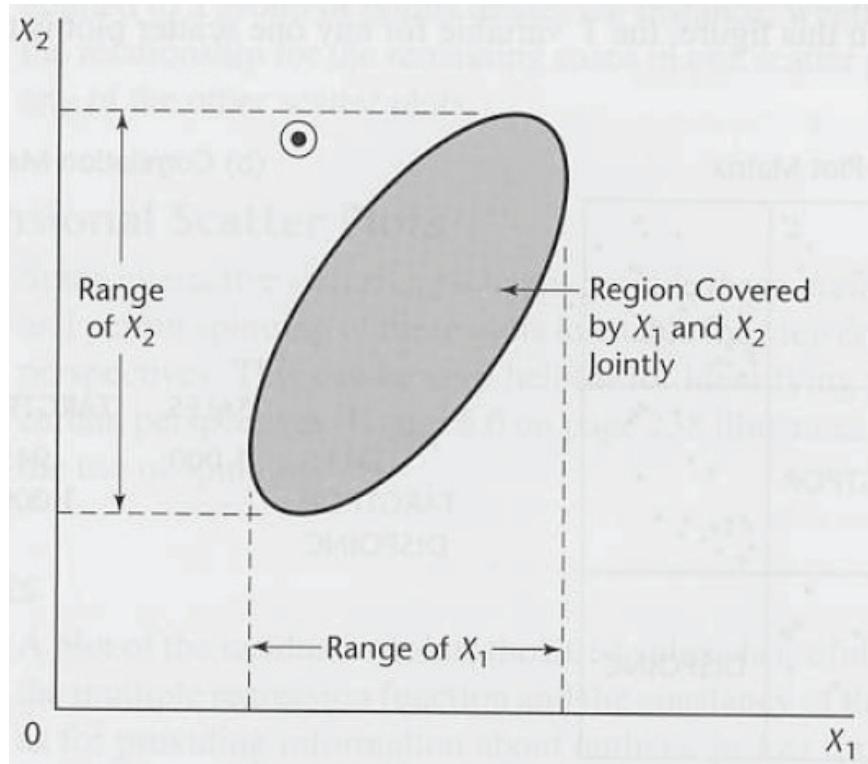
$residual.scale
[1] 6.441073

```

### 7.14.1 Another Warning About Extrapolation

Given the previous inferences using  $\mathbf{C} = \mathbf{x}^T$ , we warn again about extrapolation.

WARNING: Beware of extrapolation beyond the data used to fit the model!



## 7.15 Summary

We began with a warning about the difficulty of determining the frequentist properties (e.g., p-values, confidence interval coverage rates) associated with multiple comparisons (tests/intervals) that are suggested by an exploration of the data, as opposed to an *a priori* (pre-specified) set of comparisons, for which some control of frequentist properties exist. Though we used the flower randomized experiment in an exploratory analysis, I suspect the actual, original analysis may have been more confirmatory (with pre-specified inferences), or, at least, it seems like it should have been more confirmatory than our analysis.

Various plots, usually of residuals or fitted values, are useful for model building along with tests to add/omit variables. Overall F-test results, along

with  $R^2$  or adjusted- $R^2$  ( $R_a^2$ ) are often presented as brief summaries of model goodness, but typically do not serve alone but are accompanied by various other tests/intervals about other quantities of interest, typically some general linear combination  $\mathbf{C}\boldsymbol{\beta}$ , including specific effects,  $\beta_j$ , or about the regression function at some specified covariate value  $E(Y | \mathbf{x})$ . When we “estimate” the random variable  $Y | \mathbf{x}$  (instead of its mean  $E(Y | \mathbf{x})$ ), we say that we are “predicting,” instead of “estimating,” and machine learning does not use such distinction and typically uses “prediction” to refer to either one (usually estimating  $E(Y | \mathbf{x})$ ).

We looked at computing tests and intervals using various “hand” computations to illuminate the details behind the output convenient R functions such as `lm`, `anova`, `confint`, `predict`, `glh.test`, `estimable`, which we will see again when we cover ANOVA.

# Lecture 8

## Bayesian Linear Model

### Contents

---

|       |  |     |
|-------|--|-----|
| 8.1   | Introduction   | 275 |
| 8.2   | Distributions: Data, Prior & Posterior   | 275 |
| 8.2.1 | Data Distribution  | 276 |
| 8.2.2 | Bayes Theorem: Data, Prior, Joint, Marginal & Posterior                        | 276 |
| 8.3   | Summary So Far   | 283 |
| 8.4   | Linear Model   | 284 |
| 8.5   | Conjugate Prior  | 285 |
| 8.5.1 | Posterior  | 286 |
| 8.5.2 | Marginal Posterior for $\beta$ is a $t$  | 288 |
| 8.5.3 | Posterior Predictive is a $t$  | 290 |
| 8.5.4 | Remarks  | 291 |
| 8.6   | A Common Improper Prior  | 292 |
| 8.6.1 | Posterior  | 293 |
| 8.6.2 | Marginal Posterior for $\beta$ is a Familiar $t$                               | 293 |
| 8.6.3 | Posterior Predictive is a Familiar $t$   | 294 |
| 8.7   | STAT 101 Redux a la Bayes  | 295 |
| 8.7.1 | $t$ -based Intervals for $\beta_j$   | 296 |
| 8.7.2 | $t$ -based Test for $\beta_j$  | 297 |
| 8.7.3 | $t$ -based Intervals for $E(Y   \mathbf{x}) = \mathbf{x}^t \boldsymbol{\beta}$ | 299 |
| 8.7.4 | $t$ -based Prediction Intervals for $Y   \mathbf{x}$                           | 300 |
| 8.8   | Example  | 301 |
| 8.8.1 | Frequentist R Summary  | 301 |
| 8.8.2 | Bayesian Summary   | 302 |

|  |            |
|--|------------|
| <b>8.9 A Common Independence Prior . . . . .</b>                             | <b>307</b> |
| 8.9.1 Full Conditional Posterior Distributions . . . . .                     | 308        |
| <b>8.10 2-Stage Gibbs Sampling . . . . .</b>                                 | <b>309</b> |
| <b>8.11 Example . . . . .</b>  | <b>309</b> |
| 8.11.1 Eliciting a Prior . . . . .   | 310        |
| <b>8.12 Hamiltonian Monte Carlo in Stan . . . . .</b>                        | <b>314</b> |
| 8.12.1 Functions Block . . . . .   | 314        |
| 8.12.2 Data Block . . . . .  | 314        |
| 8.12.3 Transformed Data Block . . . . .                                      | 315        |
| 8.12.4 Parameters Block . . . . .  | 315        |
| 8.12.5 Transformed Parameters Block . . . . .                                | 315        |
| 8.12.6 Model Block . . . . .   | 316        |
| 8.12.7 Generated Quantities Block . . . . .                                  | 316        |
| 8.12.8 Altogether for Stan . . . . .   | 316        |
| 8.12.9 Translate Stan to C++ with <code>stanc</code> . . . . .               | 318        |
| 8.12.10 Make an Executable Stan Model with <code>stan_model</code> . . . . . | 318        |
| 8.12.11 Data List for Stan . . . . .   | 319        |
| 8.12.12 List of Initial Value Lists for Stan . . . . .                       | 319        |
| 8.12.13 Executing a Stan Model with <code>sampling</code> . . . . .          | 320        |
| 8.12.14 Posterior Summaries with <code>coda</code> . . . . .                 | 320        |
| <b>8.13 Example Summary . . . . .</b>  | <b>325</b> |
| <b>8.14 Other Priors . . . . .</b>   | <b>327</b> |

---

**Main Objectives:**

- Bayesian linear model
- Bayes theorem, prior distribution, posterior distribution, prior predictive distribution, posterior predictive distribution, conjugate prior, improper prior, full conditional distribution.
- Shrinkage
- Gibbs sampling

- Stan and rstan

$\mathcal{O}$

***Additional Reading:***

[Bis06, §3.3], [Mur12, §4.4], [Wak13, §3.2, 3.4, 3.7, 3.8, 3.12, 4.4, 4.11, 5.12, 5.13] \_\_\_\_\_  $\mathcal{R}$

## 8.1 Introduction

Now is a good time to review §5.2 Distributions: Joint, Marginal & Conditional. The next section parallels that section, but uses different notation more typical to a Bayesian context.

## 8.2 Distributions: Data, Prior & Posterior

- The distributions in this section's heading correspond to a conditional, a marginal & a conditional, respectively, as we discussed generally in §5.2.
- From the multiplication (i.e., product) rule (Definition 5.6), along with conditional independencies in a chosen model, we obtain a joint distribution from our data and prior distributions—this process is often referred to as **hierarchical modeling or multi-level modeling**.
- From this joint distribution, we obtain, in principle, other distributions.
- In particular, we use **Bayes Theorem** to obtain the conditional distribution known as the **posterior distribution** that is the target of Bayesian analysis.
- We will also meet the **prior predictive** (marginal) distribution and the **posterior predictive** (conditional) distribution.
- Most of the material in this section is found in [Wak13, Sec. 3.2].
- (We do not formally pursue Bayesian decision theory in which we add another element—a loss function—and decisions or actions or estimates are typically the the target. However, we may point out typical estimates and their corresponding loss function as we go: mean and squared error loss, median and absolute error loss).

### 8.2.1 Data Distribution

- We will sometimes use the term **data distribution** to refer to our model for our data.
- Let  $\mathbf{Y}$  be a random vector of **observables** (responses/outputs) whose distribution depends on a vector of parameters, generically denoted as  $\boldsymbol{\theta}$  for the moment.
- In anticipation of  $\boldsymbol{\theta}$  being formally considered as a random variable in a Bayesian framework, we may write the random vector as  $\mathbf{Y} | \boldsymbol{\theta}$  and denote its (conditional) distribution as

$$[\mathbf{y} | \boldsymbol{\theta}].$$

- E.g., classical linear model, in class
- It's conventional, among Bayesian statisticians, to explicitly condition on random variables ( $\boldsymbol{\theta}$  is now an rv...see below) and to omit conditioning on fixed quantities, like covariate/input  $\mathbf{x}$ , in a linear model, but this convention is not universally followed, particularly outside of the Bayesian statistical literature.
- Recall, if we view the data distribution as a function of the parameters, data held fixed at their observed values, we get the **likelihood** (function) of the data (§6.4). Nothing new here.

### 8.2.2 Bayes Theorem: Data, Prior, Joint, Marginal & Posterior

- For a Bayesian analysis, we consider other distributions, aside from the data (conditional) distribution, to complete a statistical model specification.
- Until now, we have conceptualized the unknown parameters (e.g.,  $\beta$  or  $\sigma^2$ ) in our data distribution as **unknown but fixed** quantities to

be estimated, with **uncertainty** in estimation following from **sampling distribution theory** (in simple cases, e.g. [Wak13, p. 29 & Ch. 5] and much of our previous notes involving  $t$  and  $F$  distributions) or from **asymptotic theory** (e.g., much of [Wak13, Ch. 2]).

- In a Bayesian analysis, we take a different tack, and we characterize our **uncertainty** about unknown parameters more directly via a **prior distribution** (a marginal distribution) and a **posterior distribution** (a conditional distribution) for the parameter itself (rather than with the distribution of an estimator of the fixed parameter, as in our  $t$  and  $F$  material so far). See [Wak13, Ch. 3].
- In other words, we may still conceptualize the parameters as fixed, but our accounting of uncertainty of the fixed quantities is formalized by probability distributions for parameters. That is, we treat the parameters as random as a way to characterize our uncertainty about them.
- In a Bayesian analysis, the goal is often to obtain the conditional distribution

$$[\boldsymbol{\theta} | \mathbf{y}].$$

- Imagine having the distribution of your unknown quantities—that accounts for the information in your data (via the likelihood as we will see) and for prior information (via the prior as we will see)—for use in inferring about the unknown quantities given what you have observed. This seems like an agreeable object to have for making inference.
- Also, as we considered briefly in the frequentist linear model (§6.7.9), we may want to predict an as yet unobserved response/output (vector),  $\mathbf{y}$ , with less emphasis on parameter  $\boldsymbol{\theta}$ .
- We might simply include such unknowns to be predicted in  $\boldsymbol{\theta}$ , but we often give explicit recognition to these unknown quantities in the

form of the **posterior predictive** distribution,

$$[\mathbf{z} | \mathbf{y}],$$

where we use  $\mathbf{z}$  to denote unobserved observables (e.g., responses/outputs we want to predict), and continue to use  $\mathbf{y}$  to denote observed observables/responses/outputs.

- Again, this seems to be a nice object to have for inferring about unknown quantities, given what you have observed.
- (In machine learning, we often hear of “prediction” as the goal, but, usually, the target is  $E(y | \mathbf{x})$ , not  $Y | \mathbf{x}$ , i.e., what machine learners call “prediction” is more often than not what we call estimation of the regression function.)
- More generally, we might consider our goal to be the distribution of

$$[unknowns | knowns],$$

which, given our discussion so far, we may denote as  $[\mathbf{z}, \boldsymbol{\theta} | \mathbf{y}]$  (a **joint posterior** of sorts).

We work our way to  $[\boldsymbol{\theta} | \mathbf{y}]$ ,  $[\mathbf{z} | \mathbf{y}]$  and  $[\mathbf{z}, \boldsymbol{\theta} | \mathbf{y}]$ , first, generically, in what follows, then in terms of the linear model, a bit later.

Recognizing the (joint) posterior

$$[\mathbf{z}, \boldsymbol{\theta} | \mathbf{y}]$$

as a joint (conditional) distribution, we can, in principle, obtain the posterior predictive,

$$[\mathbf{z} | \mathbf{y}],$$

a particular marginal (conditional) distribution, by integrating (or summing) over  $\boldsymbol{\theta}$  (see the Definition 5.4 of a marginal distribution). But, we’ll

ignore unknown responses/inputs,  $\mathbf{z}$ , for the moment, and focus on  $\boldsymbol{\theta}$ , first.

- If we specify a (marginal) **prior distribution**

$$[\boldsymbol{\theta}],$$

then, with the data (conditional) distribution, we can **build a joint distribution using the multiplication rule** (Definition 5.6)

$$[\mathbf{y}, \boldsymbol{\theta}] = [\mathbf{y} | \boldsymbol{\theta}] [\boldsymbol{\theta}].$$

- From this joint distribution, we can (in principle) get our desired (**posterior**) conditional distribution as

$$[\boldsymbol{\theta} | \mathbf{y}] = \frac{[\mathbf{y}, \boldsymbol{\theta}]}{[\mathbf{y}]}, \quad \text{or}$$

$$[\boldsymbol{\theta} | \mathbf{y}] = \frac{[\mathbf{y} | \boldsymbol{\theta}] [\boldsymbol{\theta}]}{[\mathbf{y}]}.$$

See [Wak13, Expr. (3.1)]. This result is known as **Bayes Theorem**, and we sometimes hear the term **inverse probability**. (Notice how close we were to Bayes Theorem in Definition 5.5.)

- We'll discuss prior/posterior distributions, in the context of linear models, soon.
- Notice that the marginal distribution (**prior predictive**),  $[\mathbf{y}]$ , is a constant wrt  $\boldsymbol{\theta}$  in the posterior distribution, thus it does not help to distinguish among different values of  $\boldsymbol{\theta}$ ; it is not a necessary part of the **kernel** ("important factor") of the posterior distribution and is part of its **normalizing constant** (or is often referred to as *the* normalizing constant [Wak13, Expr. (3.2)]).

- If we **recognize the kernel** of the posterior, great! “Hey, that factor looks like it belongs to a normal distribution for  $\boldsymbol{\theta}$  (up to a normalizing constant that does not depend on  $\boldsymbol{\theta}$ )!” In this case, we’re done, up to summarizing our posterior, at least. We don’t have to integrate (or sum; what?) to get the normalizing constant to get the posterior.
- In this case, we might write **Bayes Theorem** as

$$[\boldsymbol{\theta} | \mathbf{y}] \propto [\mathbf{y} | \boldsymbol{\theta}] [\boldsymbol{\theta}],$$

which we may often hear as ([Wak13, p. 86])

the posterior is proportional to the likelihood times the prior,

though we may also say (but almost never hear) “the posterior is proportional to the data distribution times the prior.”

- But, for all but the simplest situations, **we may not recognize the kernel** of the posterior. We may think then that we must somehow use numerical quadrature methods to integrate (sometimes very high dimensional) or to otherwise numerically approximate integrals (see, e.g., [Wak13, §3.7]) in order to obtain  $[\mathbf{y}]$  for use in our original statement of Bayes Theorem. But...
- We tend to focus more on Markov chain Monte Carlo (**MCMC**) methods, which effectively get around such integration. (We may view (MC)MC methods as integration methods, but, as typically implemented, (MC)MC tend to obscure this view.)
- That is, most MCMC methods do not require the normalizing constant,  $[\mathbf{y}]$ , to effectively reproduce the posterior  $[\boldsymbol{\theta} | \mathbf{y}]$ . (Poor normalizing constant, almost no one needs you...)
- Thus, again, we have the tendency to express **Bayes Theorem** as

$$[\boldsymbol{\theta} | \mathbf{y}] \propto [\mathbf{y} | \boldsymbol{\theta}] [\boldsymbol{\theta}].$$

- Classical Bayesian linear models are relatively simple, and we often recognize (conditional) kernels to give us the posterior or give us full conditional posterior distributions which are a step toward the posterior.

Now let's move toward the **posterior predictive**,  $[z | y]$ , in the case that we want to infer about unobserved responses/outputs  $z$  rather than or in addition to parameters,  $\theta$ .

- The marginal distribution (normalizing constant of the posterior, above),

$$[y] = \int [y | \theta] [\theta] d\theta,$$

(or summing) is sometimes referred to as the **prior predictive distribution**, as noted above, because it is (in principle) the distribution we would use to infer about (predict) our data  $y$ , before we observe it, after integrating/summing out, i.e., accounting for the uncertainty of the quantities we don't know, the unknown parameter  $\theta$ .

- It is **sometimes used to assess** models (likelihood and prior) by comparing it to observed data. Intuitively, if we plug in our actual, observed data,  $y$ , into  $[y]$ , and we get an unusually small value, then this suggests that we have not done a good job, a priori, of predicting our data, suggesting remodeling of our likelihood or prior or both. Or, similarly, we can compare (plot) observed data to fake data generated from  $[y]$ .
- Also, the prior predictive distribution is sometimes used in an **empirical Bayes** analysis to specify a prior distribution; sometimes called **type-II maximum likelihood** ([Ber85]). While we have integrated/summed over  $\theta$  to obtain  $[y]$ ,  $[y]$  still contains the ("hyper") parameters of the prior  $[\theta]$ , and these need to be given particular values. Thus, we view  $[y]$  as a likelihood to be maximized, the resulting

estimates of the hyperparameters plugged into  $[\boldsymbol{\theta}]$ , which is then used in a subsequent (approximate, technically not fully Bayesian) analysis. Incidentally, machine learners refer to this approach to prior specification as **evidence approximation** ([Bis06, §3.5]).

More forwardly, of course, we may want to make predictions about unobserved outputs using observations that we have observed.

- This goal is embodied (in principle) in the **(posterior) predictive distribution**,  $[z | \mathbf{y}]$ .
- As mentioned, in principle, we can obtain this by integrating/summing over the joint posterior distribution  $[z, \boldsymbol{\theta} | \mathbf{y}]$ , which we may factor (by the multiplication rule, Definition 5.6) as

$$[z, \boldsymbol{\theta} | \mathbf{y}] = [z | \mathbf{y}, \boldsymbol{\theta}][\boldsymbol{\theta} | \mathbf{y}].$$

- Thus, in principle,

$$[z | \mathbf{y}] = \int [z | \mathbf{y}, \boldsymbol{\theta}][\boldsymbol{\theta} | \mathbf{y}] d\boldsymbol{\theta},$$

where  $[z | \mathbf{y}, \boldsymbol{\theta}]$  is the distribution of unobserved data (unobserved response/outputs)  $z$  conditional on (observed)  $\mathbf{y}$ , which we can write as  $[z | \mathbf{y}, \boldsymbol{\theta}] = [z | \boldsymbol{\theta}]$ , *if*  $z$  and  $\mathbf{y}$  are independent; as in, e.g., the classical linear model. See [Wak13, Expr. (3.9)].

- As previously discussed in the case of the posterior,  $[\boldsymbol{\theta} | \mathbf{y}]$ , if we recognize the kernel of the posterior predictive, great; we have our answer (up to summarization anyway). (“Hey, that’s a t distribution!”)
- But, generally, as with  $[\boldsymbol{\theta} | \mathbf{y}]$ , for all but the simplest models, we will not recognize the kernel of the posterior predictive, and we may think that we have to evaluate the integral. But, using MCMC procedures, practically speaking, we do not have to evaluate the integral (well,

technically, we more/less evaluate it via MC(MC) integration, but this is sort of hidden).

- Again, **linear models are relatively simple** in that we often recognize kernels, but, still, we may use MCMC anyway as a convenient way to get posterior summaries, mean, variance, intervals, etc.
- Notice, because  $[z | y, \theta]$  (or, with independence,  $[z | \theta]$ ) typically follows from the specification of the data model (e.g., regression model), this gives us a relatively easy way to obtain (samples from) the posterior predictive. That is, if we can obtain samples of  $\theta$  from the (marginal conditional!) posterior  $[\theta | y]$  (from MCMC for example), and if we know the conditional  $[z | y, \theta]$  (again, it should be obvious from our data model!), then we can generate  $z | \theta, y$  so that we obtain samples of  $(z, \theta | y)$  from the joint posterior...one  $\theta | y$  then one  $z | \theta, y$ , another  $\theta | y$  then another  $z | \theta, y$ , etc. If you want a sample from the (marginal posterior),  $[z | y]$  (our posterior predictive), then just ignore the sample of  $\theta | y$  values (and vice-versa)! This process sometimes called **composition sampling**; see [Wak13, §3.8.4]

### 8.3 Summary So Far

To summarize/reiterate, we have (i.e., we specify) the data distribution (likelihood) and the prior, which, in principle, give the (joint) posterior via Bayes theorem, up to a normalizing constant anyway. Given our discussion of the **kernel** as the “important part”, we might think we are done if we recognize the kernel, because, in this case, we can avoid having to compute the integral/sum to get the normalizing constant. But, again, this requires that we recognize the form of the kernel to correspond to a known distribution, perhaps, e.g., normal, normal-gamma or t, from which we could compute easier-to-grasp summarizing quantities for our posterior, like means, medians, modes, quantiles, intervals, etc., using

ordinary methods implemented in many softwares.

When we fail to recognize kernels, we avail ourselves of other methods, such as MCMC, which usually does not require the kernel.

## 8.4 Linear Model

- Recall our linear model from previous chapters of our lecture notes (e.g., §5.3),

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma),$$

where, for much of what we will do,  $\Sigma = \sigma^2 \mathbf{I}$ .

- Thus, following our generic Bayesian discussion, above, we write our **data distribution** in the current context as

$$[\mathbf{y} | \boldsymbol{\beta}, \sigma^2].$$

- In the classic Bayesian linear model, as in the frequentist version,  $\mathbf{x}$  will be considered known (see §5.3.1), and we may not use notation to explicitly indicate conditioning on  $\mathbf{x}$  or  $\mathbf{X}$  (matrix). This assumes a factorization of the prior analogous to that for  $[\mathbf{y}, \mathbf{x}]$ , discussed in §5.3.1, so that we do not have to consider the prior for the parameters of the covariate distribution, but we skip further discussion of this.
- Continuing to follow the above, generic Bayesian development, we must specify a **prior distribution** for the parameters, which, in the current context, we denote as

$$[\boldsymbol{\beta}, \sigma^2].$$

- We will discuss common prior distributional specifications and corresponding **posterior distribution**, which, in our current context, we denote as

$$[\boldsymbol{\beta}, \sigma^2 | \mathbf{y}].$$

- And, we well will discuss the corresponding **posterior predictive** distribution, which we denote now as

$$[\mathbf{y}^* | \mathbf{y}],$$

where  $\mathbf{y}^*$  denotes a vector of  $n^*$  unobserved outputs, with associated inputs,  $\mathbf{x}^*$ , that we would like to predict.

- ( $\mathbf{y}^*$  corresponds to  $\mathbf{z}$  in the generic discussion, above.)
- Again, we tend to suppress notation on covariates,  $\mathbf{x}$ ,  $\mathbf{X}$  (matrix) or  $\mathbf{x}^*$ .
- In our previous, frequentist material, we didn't put stars \* on values to be predicted or their associated covariates, but perhaps we should have for consistency in notation.

## 8.5 Conjugate Prior

- The **conjugate prior distribution** for the mean parameter,  $\boldsymbol{\beta}$ , and variance parameter,  $\sigma^2$ , considered collectively in the normal linear model, is the product (using the multiplication rule) of a **conditional normal distribution** and a **scaled inverse-chi-square**,

$$\begin{aligned} [\boldsymbol{\beta}, \sigma^2] &= [\boldsymbol{\beta} | \sigma^2][\sigma^2] \\ &= N(\mathbf{m}_0, \sigma^2 \boldsymbol{\Sigma}_0) \times \text{inv-}\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

where we must specify values (or further distributions—not here) for the prior (hyper)parameters,

- $\mathbf{m}_0$  = conditional prior **mean**,
- $\sigma^2 \boldsymbol{\Sigma}_0$  = conditional prior **variance**,
- $\nu_0$  = marginal prior **degrees of freedom**
- $\sigma_0^2$  = marginal prior (squared) **scale**.

- See [Wak13, §3.7.1 & p. 223] for a brief discussion of conjugacy. (Incidentally,  $\text{inv-}\chi^2(\nu_0, \sigma_0^2)$  is the same as  $\text{inv-gamma}(a = \nu_0/2, b = \sigma_0^2\nu_0/2)$ .)
- Shortly, we will see more intuition for why  $\nu_0$  and  $\sigma_0^2$  are called degrees of freedom and (squared) scale, respectively, and hence why  $\nu_0\sigma_0^2$ , may be seen as a **prior sum-of-squares**.
- We will attempt to follow the **distribution parameterizations** of [GCS<sup>+</sup>14, Appendix A], unless otherwise indicated.
- A **conjugate prior distribution** for the parameter of a data distribution is such that the posterior distribution of that parameter is of the **same distributional form** as the prior distribution for that parameter ([Rob01, Def. 3.3.1]). The data distribution and the conjugate prior are often said to form a **conjugate pair** of distributions.
- Thus, given our conjugate prior above, e.g., we should expect the posterior (see below) to also be the product of a **conditional normal distribution** and a **scaled inverse-chi-square** (with different hyperparameters that are somehow updated to incorporate our observed data, of course).
- Conjugacy is not only convenient, but has **theoretical justification**. For certain families of data distributions, including ours, here, mixtures of conjugate distributions are also conjugate, and we can get arbitrarily close to many distributions with such mixtures ([Rob01, Lemma 3.4.2 & Theorem 3.4.3]).

### 8.5.1 Posterior

- According to **Bayes theorem**, we have

$$[\boldsymbol{\beta}, \sigma^2 | \mathbf{y}] \propto [\mathbf{y} | \boldsymbol{\beta}, \sigma^2][\boldsymbol{\beta} | \sigma^2][\sigma^2],$$

which, by conjugacy (and other standard results), can be shown to give the **posterior**

$$\begin{aligned} [\boldsymbol{\beta}, \sigma^2 | \mathbf{y}] &= [\boldsymbol{\beta} | \sigma^2, \mathbf{y}][\sigma^2 | \mathbf{y}] \\ &= N(\widehat{\mathbf{m}}, \sigma^2 \widehat{\Sigma}) \times \text{inv-}\chi^2(\widehat{\nu}, \widehat{\sigma}^2) \end{aligned}$$

where

$$\begin{aligned} \widehat{\mathbf{m}} &= (\Sigma_0^{-1} + (\mathbf{X}^t \mathbf{X}))^{-1} (\Sigma_0^{-1} \mathbf{m}_0 + (\mathbf{X}^t \mathbf{X}) \widehat{\boldsymbol{\beta}}) \quad \text{cond. post. mean,} \\ \sigma^2 \widehat{\Sigma} &= \sigma^2 (\Sigma_0^{-1} + (\mathbf{X}^t \mathbf{X}))^{-1} \quad \text{cond. post. variance,} \\ \widehat{\nu} &= \nu_0 + n \quad \text{post. df} \\ \widehat{\sigma}^2 &= (1/\widehat{\nu})(\nu_0 \sigma_0^2 + (\mathbf{y} - \mathbf{X} \widehat{\boldsymbol{\beta}})^t (\mathbf{y} - \mathbf{X} \widehat{\boldsymbol{\beta}}) + \\ &\quad (\mathbf{m}_0 - \widehat{\boldsymbol{\beta}})^t (\Sigma_0 + (\mathbf{X}^t \mathbf{X})^{-1})^{-1} (\mathbf{m}_0 - \widehat{\boldsymbol{\beta}})) \text{post. scale (squared)} \end{aligned}$$

- Thus, conjugacy simply requires that we use (well-established) **update formulas** to get from our prior to our posterior; the result of Bayes theorem has largely been obtained for us with relatively little effort (just some algebra, which you will undoubtedly find all over the Web).

- **Notice a few things:**

- The conditional normal **posterior mean** of  $\boldsymbol{\beta}$  is a weighted average of the prior mean of  $\boldsymbol{\beta}$  and a data-based (unbiased) estimate of  $\boldsymbol{\beta}$ .
- The conditional normal **posterior variance** of  $\boldsymbol{\beta}$  is the inverse of the sum of its prior precision (inverse of variance) and the precision of a data-based estimate.
- The marginal **posterior df** is a sum of the prior degrees of freedom and sample size.
- The marginal **posterior sum-of-squares** is a sum of the prior sum-of-squares (now we see), the usual error sum-of-squares (SSE) and a sum-of-squares that is a sort of discrepancy between the prior and data-based means of  $\boldsymbol{\beta}$ .

- As the (conditional) prior variance “increases,” the discrepancy between prior (conditional) mean and data-based mean is down-weighted, making the posterior scale (squared) smaller hence making the posterior variance (and mean) of  $\sigma^2$  smaller. (You need know the mean and variance of a scaled inv- $\chi^2$  for this comment to make sense ([GCS<sup>+</sup>14, Appendix A]).)
- As prior df increases the posterior scale (squared) looks more like the prior scale (squared) hence the posterior mean (and variance) of  $\sigma^2$  looks more like the prior scale (squared). (You need know the mean and variance of a scaled inv- $\chi^2$  for this to make sense ([GCS<sup>+</sup>14, Appendix A]).)

### 8.5.2 Marginal Posterior for $\beta$ is a $t$

- The **marginal posterior**,  $[\beta | \mathbf{y}]$ , is what we would use to infer about  $\beta$ . In some (rough) sense, it’s the Bayesian counterpart to the normal/ $\chi^2$ /t/F results in chapter 6 for inferring about  $\beta$ .
- We will use the following **result about a t distribution** to get  $[\beta | \mathbf{y}]$  (and to get the posterior predictive,  $[\mathbf{y}^* | \mathbf{y}]$ , shortly thereafter).
- The (generic, not sample size)  $n$  dimensional  $t$  distribution (pdf),

$$t_n(\nu, \boldsymbol{\mu}, \sigma_0^2 \boldsymbol{\Sigma}),$$

is often defined as a **scale mixture** of a conditional normal with an inv- $\chi^2$  (scaled inverse chi-square) mixing distribution (alternatively and equivalently, an inv-gamma mixing distribution, but I prefer the parameterization of an inv- $\chi^2$ ).

- In other words, the  $t$  pdf is often defined as the marginal distribution,

$$[\boldsymbol{\theta}] = t_n(\nu, \boldsymbol{\mu}, \sigma_0^2 \boldsymbol{\Sigma}),$$

where  $(\boldsymbol{\theta}^t, \sigma^2)^t$  has joint distribution

$$[\boldsymbol{\theta}, \sigma^2] = [\boldsymbol{\theta} | \sigma^2][\sigma^2]$$

defined by the conditional normal

$$\boldsymbol{\theta} | \sigma^2 \sim N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$$

and marginal inv- $\chi^2$

$$\sigma^2 \sim \text{inv-}\chi^2(\nu, \sigma_0^2).$$

- This means, any time we see this form for a joint distribution, we can immediately write down the marginal

$$[\boldsymbol{\theta}] = t_n(\nu, \boldsymbol{\mu}, \sigma_0^2 \boldsymbol{\Sigma}).$$

- Incidentally, this is the multivariate, non-centered, scaled (*not* non-central) version of the univariate standard  $t$  of “STAT 101,”, i.e., of the **Student’s t**. In other words, if you multiplied a standard (Student’s)  $t$  (with  $df = \nu$ ) by a scalar,  $\sigma_0$ , then added  $\mu$ , you would have a variable distributed as

$$t \sim t_1(\nu, \mu, \sigma_0^2),$$

where  $E(t) = \mu$  ( $\nu > 1$ ) and  $\text{Var}(t) = \frac{\nu}{\nu-2} \sigma^2 z$  ( $\nu > 2$ ).

- (NOTE: [GCS<sup>+</sup>14, Appendix A] absorbs  $\sigma_0^2$  into  $\boldsymbol{\Sigma}$ , and [Wak13, Appendix D] uses different symbols in a different order.)
- Thus, we can use this result to get our marginal ( $t$ ) posterior (right!?),

$$[\boldsymbol{\beta} | \mathbf{y}] = t_p(\widehat{\nu}, \widehat{\boldsymbol{m}}, \widehat{\sigma}^2 \widehat{\boldsymbol{\Sigma}}),$$

where all updated posterior parameters are as defined above.

### 8.5.3 Posterior Predictive is a $t$

- What's the **posterior predictive**,  $[y^* | \mathbf{y}]$ ?
- Our model for  $n^*$  unobserved, “future” responses/outputs, which we denote as  $\mathbf{y}^*$ , to distinguish them from observed inputs,  $\mathbf{y}$ , follow our same data model,

$$\mathbf{y}^* \sim N(\mathbf{X}^* \boldsymbol{\beta}, \sigma^2 \mathbf{I}),$$

independent of  $\mathbf{y}$ .

- Thus, we have a **joint data (unobserved and observed) distribution**

$$[y^*, \mathbf{y} | \boldsymbol{\beta}, \sigma^2] = [y^* | \mathbf{y}, \boldsymbol{\beta}, \sigma^2][\mathbf{y} | \boldsymbol{\beta}, \sigma^2] = [y^* | \boldsymbol{\beta}, \sigma^2][\mathbf{y} | \boldsymbol{\beta}, \sigma^2],$$

which we can use in the numerator of Bayes Theorem.

- By the multiplication rule, we can write the **joint posterior** as

$$[y^*, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}] = [y^*, \boldsymbol{\beta} | \sigma^2, \mathbf{y}][\sigma^2 | \mathbf{y}],$$

and, it does not take terribly much algebra to show that, as before,

$$[y^*, \boldsymbol{\beta} | \sigma^2, \mathbf{y}]$$

is a **conditional normal** distribution, and

$$[\sigma^2 | \mathbf{y}]$$

is the same **marginal inv- $\chi^2$** , above.

- ( $\mathbf{y}^*$  and  $\boldsymbol{\beta}$  collectively now play the role of  $\boldsymbol{\beta}$ , above, or of the generic  $\boldsymbol{\theta}$  in the above t result).
- Before applying the above t result, we “integrate out” (not!)  $\boldsymbol{\beta}$  to get a (marginal) conditional normal,

$$[y^* | \sigma^2, \mathbf{y}],$$

thus getting a marginal posterior that is the product of a conditional normal posterior and the same inv- $\chi^2$ ,

$$[\mathbf{y}^*, \sigma^2 | \mathbf{y}] = [\mathbf{y}^* | \sigma^2, \mathbf{y}][\sigma^2 | \mathbf{y}],$$

where, in particular, (skipping some details)

$$\begin{aligned} [\mathbf{y}^* | \sigma^2, \mathbf{y}] &= N(\hat{\boldsymbol{\mu}}^*, \sigma^2 \hat{\boldsymbol{\Sigma}}^*), \\ \hat{\boldsymbol{\mu}}^* &= \mathbf{X}^*(\boldsymbol{\Sigma}_0^{-1} + (\mathbf{X}^t \mathbf{X}))^{-1}(\boldsymbol{\Sigma}_0^{-1} \mathbf{m}_0 + (\mathbf{X}^t \mathbf{X}) \hat{\boldsymbol{\beta}}), \\ &= \mathbf{X}^* \hat{\mathbf{m}} \\ \hat{\boldsymbol{\Sigma}}^* &= (\mathbf{I} + \mathbf{X}^*(\boldsymbol{\Sigma}_0^{-1} + (\mathbf{X}^t \mathbf{X}))^{-1} \mathbf{X}^{*t}), \\ &= (\mathbf{I} + \mathbf{X}^* \hat{\boldsymbol{\Sigma}} \mathbf{X}^{*t}), \end{aligned}$$

and  $[\sigma^2 | \mathbf{y}]$  is the same inv- $\chi^2$  as above with the same df and scale parameters as defined above.

- Thus, we can use the above t result to write down the desired **posterior predictive** distribution as

$$[\mathbf{y}^* | \mathbf{y}] = t_{n^*}(\hat{\nu}, \hat{\boldsymbol{\mu}}^*, \hat{\sigma}^2 \hat{\boldsymbol{\Sigma}}^*),$$

where all posterior parameters have been defined above.

#### 8.5.4 Remarks

- Alas, the above conjugate normal×inv- $\chi^2$  prior for linear models may seem **unrealistic** in the sense that  $\boldsymbol{\beta}$  becomes increasingly concentrated (disperse) around its prior mean,  $\mathbf{m}_0$ , as  $\sigma^2$  becomes smaller (larger). Or, this just may seem like an unnatural way to express prior information about  $\boldsymbol{\beta}$ .
- **Zellner's G-prior** fits into the current discussion (and that of the next section) by taking  $\boldsymbol{\Sigma}_0 = g(\mathbf{X}^t \mathbf{X})^{-1}$  and specifying a value (or prior) for  $g > 0$  (and setting  $[\sigma^2] \propto \sigma^{-2}$ ; next section), which seems

to alleviate the previous item's concern (and we seem to have a more informative prior on  $\beta$  in the sense of its using its (known covariate) data-based covariance structure  $(\mathbf{X}^t \mathbf{X})^{-1}$ ). Zellner's G-prior is traditionally used in Bayesian variable selection. We may not get there. If not, I'll cover it in INF 504.

- Or, we may consider other priors. But, generally speaking, we do not get (full) conjugacy, but, instead, conditional conjugacy or perhaps no conjugacy (unusual for typical normal linear models).
- In the next section, we consider a popular **non-informative, improper prior**, that results in effectively the same form of posterior and posterior predictive (and our work above is not for naught!).

## 8.6 A Common Improper Prior

- A common **improper prior** distribution for the normal linear model is ([Wak13, Expr. (5.42)])

$$[\beta, \sigma^2] \propto \sigma^{-2}.$$

- Incidentally, this is the product of **Jeffreys' prior** for  $\beta$  ( $\propto 1$ ) and **Jeffreys' prior** for  $\sigma^2$  (but, perhaps confusingly, is *not* Jeffreys' prior for (jointly)  $(\beta, \sigma^2)$ !...)
- An **improper prior** is one that does not integrate (or sum) to a finite value, thus it cannot be normalized to integrate (or sum) to 1.
- It's the **propriety of the posterior** that counts.
- Loosely, this improper prior can be viewed as the previous section's proper conjugate prior with df  $\nu_0 = 0$  and prior precision (almost)  $\Sigma_0^{-1} = \mathbf{0}$ . (If you plugged these values into the pdf, you'd see.)

- It's **not a conjugate prior** (unless we consider the improper case as member of the normal-inv- $\chi^2$  family...).

### 8.6.1 Posterior

- In short, we have the posterior

$$\begin{aligned} [\boldsymbol{\beta}, \sigma^2 | \mathbf{y}] &= [\boldsymbol{\beta} | \sigma^2, \mathbf{y}] [\sigma^2 | \mathbf{y}] \\ &= N(\widehat{\mathbf{m}}, \sigma^2 \widehat{\Sigma}) \times \text{inv-}\chi^2(\widehat{\nu}, \widehat{\sigma}^2) \end{aligned}$$

where

$$\begin{aligned} \widehat{\mathbf{m}} &= \widehat{\boldsymbol{\beta}} \quad \text{cond. post. mean,} \\ \sigma^2 \widehat{\Sigma} &= \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1} \quad \text{cond. post. variance,} \\ \widehat{\nu} &= n - p \quad \text{degrees of freedom} \\ \widehat{\sigma}^2 &= (1/\widehat{\nu})(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^t(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \quad \text{post. scale squared} \end{aligned}$$

NOTE: now, we must have  $n > p$  and  $\mathbf{X}$  must be full rank, neither of which were strictly necessary in the conjugate case, above, as long as  $\boldsymbol{\Sigma}_0$  was a valid variance matrix (positive definite). Things are looking strangely familiar...

### 8.6.2 Marginal Posterior for $\boldsymbol{\beta}$ is a Familiar $t$

- Again, the posterior is of the form of the above t scale mixture result, giving **marginal posterior**

$$[\boldsymbol{\beta} | \mathbf{y}] = t_p(n - p, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2 (\mathbf{X}^t \mathbf{X})^{-1})$$

([Wak13, Expr. 5.44],  $p = k + 1$ , and his parameter arguments are in a different order).

- In particular, the MLE (and LS estimator) is the same as the (marginal) posterior mean,  $\hat{\beta}$  ( $n - p > 1$ ), and the posterior variance ( $n - p > 2$ ) is the same as  $\widehat{Var}(\hat{\beta})$ !
- In other words, each  $\beta_j$  is

$$t_1(n - p, \hat{\beta}_j, \hat{\sigma}^2(\mathbf{X}^t \mathbf{X})_{(jj)}^{-1})$$

where  $(\mathbf{X}^t \mathbf{X})_{(jj)}^{-1}$  is the  $j$ th diagonal element of  $(\mathbf{X}^t \mathbf{X})^{-1}$ .

- In other words, standardizing, we get

$$\frac{\beta_j - \hat{\beta}_j}{\sqrt{\hat{\sigma}^2(\mathbf{X}^t \mathbf{X})_{jj}^{-1}}} \sim t(n - p) \quad whaoahuh!?$$

- This should look very familiar (see Result 6.10), but is, in some sense, very different as, now, in the Bayesian context the randomness comes from  $\beta$  and not from  $\hat{\beta}$  and  $\hat{\sigma}^2$ ! (See related comment middle of [Wak13, p. 222].)
- To be sure,  $\beta$  ( $\beta_j$ ) is now random, not  $\hat{\beta}$  ( $\hat{\beta}_j$ ), which is now fixed, as is  $\hat{\sigma}^2$ .
- (Back in chapter 5, we could have stated a multi-variate  $t$  result for  $\beta$ , but we didn't.)

### 8.6.3 Posterior Predictive is a Familiar $t$

- And, in short, we have the **posterior predictive**

$$[\mathbf{y}^* | \mathbf{y}] = t_{n^*}(\hat{\nu}, \hat{\mu}^*, \hat{\sigma}^2 \hat{\Sigma}^*),$$

where (we've skipped a few details)

$$\hat{\boldsymbol{\mu}}^* = \mathbf{X}^* \hat{\boldsymbol{\beta}}, \quad (8.1)$$

$$\hat{\boldsymbol{\Sigma}}^* = (\mathbf{I} + \mathbf{X}^* (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^{*t}), \quad (8.2)$$

$$\hat{\nu} = n - p, \quad (8.3)$$

$$\hat{\sigma}^2 = (1/\hat{\nu})(\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^t (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) \quad (8.4)$$

- In other words,

$$[\mathbf{y}^* | \mathbf{y}] = t_{n*}(n - p, \mathbf{X}^* \hat{\boldsymbol{\beta}}, \hat{\sigma}^2 (\mathbf{I} + \mathbf{X}^* (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^{*t})),$$

which means  $y_i^*$  is

$$t_1(n - p, \mathbf{x}_i^{*t} \hat{\boldsymbol{\beta}}, \hat{\sigma}^2 (1 + \mathbf{x}_i^{*t} (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_i^*)).$$

- In other words, if we standardize, we get

$$\frac{y_i^* - \mathbf{x}_i^{*t} \hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2 (1 + \mathbf{x}_i^{*t} (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_i^*)}} \sim t(n - p)$$

- This, too, should look very familiar, but, while  $y_i^*$  is random, Bayesian or not,  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$  are now, again, fixed.

## 8.7 STAT 101 Redux a la Bayes

In light of the previous section's **particular improper prior** for our normal linear model, what do you think about the following results (as seen previously in §6.8)? More discussion in class.

NOTE: In this section, to streamline notation in the hope of bettering drawing the coincidence with frequentist results, we will often suppress conditioning for the Bayesian results, e.g., we use  $\beta_j$  instead of  $\beta_j | \mathbf{y}$  or use  $E(Y | \mathbf{x}) = \mathbf{x}^t \boldsymbol{\beta}$  instead of  $E(Y | \mathbf{y}, \mathbf{x}) = \mathbf{x}^t \boldsymbol{\beta} | \mathbf{y}$  or use  $Y | \mathbf{x}$  instead of  $Y | \mathbf{y}, \mathbf{x}$ .

### 8.7.1 $t$ -based Intervals for $\beta_j$

- Reiterating our **Bayesian** results, we have

$$\beta_j \sim t_1(n - p, \hat{\beta}_j, \hat{\sigma}^2(\mathbf{X}^t \mathbf{X})_{(jj)}^{-1}).$$

- Reiterating one of our **frequentist** sampling distribution results (note chapter 6), we have

$$\hat{\beta}_j \sim t_1(n - p, \beta_j, \hat{\sigma}^2(\mathbf{X}^t \mathbf{X})_{(jj)}^{-1}).$$

- In either case, we can standardize to get

$$Pr \left( t(n - p, \alpha/2) \leq \frac{\beta_j - \hat{\beta}_j}{\sqrt{\hat{\sigma}^2(\mathbf{X}^t \mathbf{X})_{jj}}} \leq t(n - p, 1 - \alpha/2) \right) = 1 - \alpha$$

(draw a picture).

- After a bit of algebra and using the symmetry of the  $t$  distribution, gives

$$Pr \left( \hat{\beta}_j - t(n - p, 1 - \alpha/2) \sqrt{\hat{\sigma}^2(\mathbf{X}^t \mathbf{X})_{jj}^{-1}} \leq \beta_j \leq \hat{\beta}_j + t(n - p, 1 - \alpha/2) \sqrt{\hat{\sigma}^2(\mathbf{X}^t \mathbf{X})_{jj}^{-1}} \right) = 1 - \alpha$$

(dimension 1, dropped from notation).

- In other words, under the particular improper prior under current consideration, the Bayesian and frequentist intervals are the same, quantitatively.
- However, the probability (“Pr”) is computed differently in these two cases, right?
- For frequentists, the randomness comes from the data via the hatted quantities,  $\hat{\beta}$  and  $\hat{\sigma}^2$ , and  $\beta_j$  is fixed (and unknown); after the observed data values are plugged in, the interval is fixed, and there is no randomness left, hence frequentists do not use “probability” to refer to their intervals, but, as we know, use the term **confidence intervals** with a **long-run relative frequency interpretation** using the notion of **hypothetical replications**.

- For Bayesians, the data and the hatted quantities are fixed, and  $\beta_j$  is random. Thus, the Bayesian interval is a fixed interval—numerically the same as the frequentist intervals for the prior under current consideration—within which a random quantity resides with some **probability**; still, we call these intervals **credible intervals**, and there is no need to resort to long-run relative frequency or hypothetical replications to interpret the intervals. Of course, this assumes that we somehow accept probability as serving as its own interpretation, perhaps as a degree of belief or degree of uncertainty about  $\beta_j$ .
- In short, some of the numerical results from our frequentist inference correspond exactly to the numerical results of a Bayesian inference under our particular improper prior, but interpretation is very different.

```
> bigbang.df<- Sleuth3::case0701
> bb.lm<- lm(Distance ~ Velocity, data=bigbang.df)
> confint(bb.lm)

              2.5 %    97.5 %
(Intercept) 0.1530719058 0.64526897
Velocity     0.0008999349 0.00184488
```

### 8.7.2 $t$ -based Test for $\beta_j$

The numerical correspondence of frequentist and Bayesian intervals (for the particular improper prior under consideration) holds for **two-sided intervals**, discussed here, and for **one-sided intervals**. However, while the correspondence holds for our usual **one-sided tests**, it does **not generally hold for two-sided tests with a point null hypothesis**; we often get very different numerical results when comparing frequentist p-values to Bayesian probabilities of null hypotheses consisting of a single value of the parameter. You can begin to see the reason for differences in such point null cases if you

consider that, for a continuous parameter,  $\theta$ ,  $Pr(\theta = \theta_0) = 0$ , but we do not pursue these testing differences further; see the Jeffreys-Lindley Paradox in [Wak13, §4.4].

- For one-sided tests, consider  $H_0 : \beta_j \leq b_0$  and  $H_a : \beta_j > b_0$ .
- For frequentists, we compute  $t_{stat} = (\hat{\beta}_j - b_0) / \widehat{SE}(\hat{\beta}_j)$  and p-value =  $Pr(t > t_{stat})$ .
- For Bayesians, we compute the **probability of the null hypothesis**,

$$\begin{aligned} Pr(H_0) &= Pr(\beta_j \leq b_0) \\ &= Pr\left(\frac{\beta_j - \hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)} \leq \frac{b_0 - \hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}\right) \\ &= Pr(t \leq -t_{stat}) = Pr(-t \geq t_{stat}) \\ &= Pr(t \leq -t_{stat}) = Pr(-t \geq t_{stat}) \\ &= Pr(t \geq t_{stat}) \end{aligned}$$

where that last equality follows by symmetry of the Student's (standard)  $t$  distribution, i.e.,  $t$  is distributed the same as  $-t$ .

- That is, for the particular improper prior under current consideration, the usual one-sided tests for  $\beta_j$  give the same numerical result, but, the interpretation is different.
- For frequentists, the p-value has a **long-run relative frequency interpretation** using the notion of **hypothetical replications**.
- For Bayesians, we can speak of the probability of the null being true.
- Despite the correspondence between the p-value and the probability of the null hypothesis being true, in the one-side case, generally, the **p-value is not the probability of the null hypothesis** (again, see [Wak13, §4.4] and [Wak13, §4.3]).

```

> ## one-sided tests
> bhat<- coef(bb.lm)
> se<- sqrt(diag(vcov(bb.lm)))
>
> ## Ho: bj <= b0, Ha: bj > b0
> b0<- 0 ## for illustration only
> tstat<- (bhat - b0)/se
>
> ## Freq: Pr(t > tstat):
> pt(tstat, df=bb.lm$df, lower.tail=FALSE)

(Intercept)      Velocity
1.401504e-03 2.303841e-06

> ## Bayes: Pr(bj <= b0) = Pr((bj - bhat)/se <= (b0-bhat)/se) = Pr(t <=
> ## -tstat) = Pr(-t >= tstat) = Pr(t >= tstat) (same)
>
> rm(bhat,se,b0,tmult)

```

### 8.7.3 $t$ -based Intervals for $\mathbf{E}(Y | \mathbf{x}) = \mathbf{x}^t \boldsymbol{\beta}$

We did not say it, above, but our **Bayesian** results lead to

$$\mathbf{x}^t \boldsymbol{\beta} | \mathbf{y} \sim t_1(n - p, \mathbf{x}^t \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2 \mathbf{x}^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}),$$

which follows from (unstated) properties of the  $t$  distribution and the fact that  $\boldsymbol{\beta} | \mathbf{y}$  is a  $t$  as shown above.

The (unstandardized) **frequentist** version is (see notes near Result 6.14 for standardized (Student's  $t$ ) version)

$$\mathbf{x}^t \widehat{\boldsymbol{\beta}} \sim t_1(n - p, \mathbf{x}^t \boldsymbol{\beta}, \widehat{\sigma}^2 \mathbf{x}^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}).$$

Thus, in the same way as above, we may now view the frequentist R output for our linear model from our (particular improper prior) Bayesian perspective.

```

> xnew<- data.frame(Velocity=800)
> predict(bigbang.lm, newdata=xnew, se.fit=TRUE, interval="confidence",
+           level=0.95)

```

```
$fit
  fit      lwr      upr
1 1.497096 1.232213 1.76198

$se.fit
[1] 0.1277242

$df
[1] 22

$residual.scale
[1] 0.4056302
```

#### 8.7.4 $t$ -based Prediction Intervals for $Y | \mathbf{x}$

In essentially the same way as above, omitting details, we get similar  $t$  results and may now view the frequentist R output for prediction intervals from our Bayesian perspective. ( $Y$  here denotes an unobserved response despite our not using \* notation as we did above.)

```
> predict(bigbang.lm, newdata=xnew, se.fit=TRUE, interval="predict",
+           level=0.95)

$fit
  fit      lwr      upr
1 1.497096 0.6151532 2.37904

$se.fit
[1] 0.1277242

$df
[1] 22

$residual.scale
[1] 0.4056302
```

## 8.8 Example

Here, we follow the Bayesian analysis of the prostate data given by [Wak13, §5.12] based on the improper prior of §8.6, in our notes, above ([Wak13, Expr. (5.42)]). Thus, as discussed, numerical results coincide with frequentist results. Frequentist LS (MLE) estimates (posterior means) and standard errors (posterior scales (which are almost posterior standard deviations)) are given in [Wak13, Tab. 5.14] ( $\hat{\sigma}^2$  is not the MLE, but MSE, the unbiased “ $n - p$ ” estimator (Definition 6.3)).

### 8.8.1 Frequentist R Summary

```
> library(lasso2, quietly=TRUE)
> data(Prostate)
> prostate.df<- Prostate
> detach(package:lasso2)
> names(prostate.df)

[1] "lcavol"   "lweight"   "age"        "lbph"      "svi"
[6] "lcp"       "gleason"   "pgg45"     "lpsa"

> summary(prostate.lm<- lm(lpsa ~ ., data=prostate.df))
```

Call:

```
lm(formula = lpsa ~ ., data = prostate.df)
```

Residuals:

| Min      | 1Q       | Median   | 3Q      | Max     |
|----------|----------|----------|---------|---------|
| -1.73316 | -0.37133 | -0.01702 | 0.41414 | 1.63811 |

Coefficients:

|             | Estimate  | Std. Error | t value | Pr(> t )     |
|-------------|-----------|------------|---------|--------------|
| (Intercept) | 0.669399  | 1.296381   | 0.516   | 0.60690      |
| lcavol      | 0.587023  | 0.087920   | 6.677   | 2.11e-09 *** |
| lweight     | 0.454461  | 0.170012   | 2.673   | 0.00896 **   |
| age         | -0.019637 | 0.011173   | -1.758  | 0.08229 .    |
| lbph        | 0.107054  | 0.058449   | 1.832   | 0.07040 .    |
| svi         | 0.766156  | 0.244309   | 3.136   | 0.00233 **   |
| lcp         | -0.105474 | 0.091013   | -1.159  | 0.24964      |

```

gleason      0.045136   0.157464   0.287   0.77506
pgg45        0.004525   0.004421   1.024   0.30885
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7084 on 88 degrees of freedom
Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16

```

### 8.8.2 Bayesian Summary

[Wak13, Fig. 5.10] displays the  $t$  posteriors of the regression function parameters (excluding  $\beta_0$ ) for the particular improper prior that we continue to consider, but the **covariates have been standardized** so that the posteriors are on a comparable scale. You may think of the standardization simply as a device to get the parameters on the same scale for plotting. However, we will use this standardization in a subsequent section of our notes to elicit an informative prior ([Wak13, §5.12]), shortly.

```

> standx<- function(x) {
+   rangex<- range(x)
+   (x - rangex[1]) / diff(rangex)
+ }
> zprostate.df<- prostate.df$lpsa
> for (k in 1:8) zprostate.df<-
+   cbind.data.frame(zprostate.df,
+                     standx(prostate.df[,k]))
> names(zprostate.df)<- c("lpsa",paste0("z.",names(prostate.df)[1:8]))
> round(head(zprostate.df, n=3), 3)

  lpsa z.lcavol z.lweight z.age z.lbph z.svi z.lcp
1 -0.431    0.148    0.106 0.237      0      0      0
2 -0.163    0.068    0.253 0.447      0      0      0
3 -0.163    0.162    0.085 0.868      0      0      0
  z.gleason z.pgg45
1      0.000    0.0
2      0.000    0.0
3      0.333    0.2

```

```
> round(tail(zprostate.df, n=3), 3)

lpsa z.lcavol z.lweight z.age z.lbph z.svi z.lcp
95 5.143    0.823    0.274 0.289  0.000      1 0.897
96 5.478    0.818    0.375 0.711  0.793      1 0.686
97 5.583    0.932    0.429 0.711  0.491      1 1.000
z.gleason z.pgg45
95     0.333    0.1
96     0.333    0.8
97     0.333    0.2
```

Compare the following output to the credible intervals in [Wak13, Fig. 5.11 (solid lines)]. We give the frequentist regression function parameter (interval) estimates for this standardized covariate analysis for comparison. These results *would* correspond to those in [Wak13, Tab. 5.14], but the analysis that produced those results uses the raw covariates, not the standardized covariates, thus results differ.

```
> ## Bayes t means (df>1) and LS/MLE:
> tpostmeans<- coef(zprostate.lm<- lm(lpsa ~ ., data=zprostate.df))
> ## Bayes t df (n-p) and freq error df:
> tpostdf<- zprostate.lm$df
> ## Bayes t squared scales:
> tpostscales2<- diag(vcov(zprostate.lm))
> ## Bayes t variances (df>2):
> tpostvars<- tpostdf / (tpostdf -2) * tpostscales2
> ## Bayes t standard deviations:
> tpostsds<- sqrt(tpostvars)
> ## Freq ses are Bayes t scales, not Bayes t sds:
> tfreqses<- sqrt(tpostscales2)
> ## Intervals:
> t975<- qt(p=1-alpha/2, df=tpostdf)
> tpost25lb<- tpostmeans - t975*sqrt(tpostscales2)
> tpost975ub<- tpostmeans + t975*sqrt(tpostscales2)
> ## Compare to (not Table 5.14) Fig 5.10, and Fig 5.11 (solid lines)
> round(cbind("pmean"=tpostmeans, "psd"=tpostsds,
+             "freqse"=tfreqses, "25lb"=tpost25lb,
+             "975ub"=tpost975ub), 4)

          pmean      psd freqse      25lb   975ub
(Intercept) 0.4214 0.2992 0.2958 -0.1664 1.0092
```

```

z.lcavol      3.0338 0.4596 0.4544  2.1308 3.9368
z.lweight     1.6964 0.6419 0.6346  0.4352 2.9575
z.age        -0.7462 0.4295 0.4246 -1.5899 0.0975
z.lbph       0.3974 0.2195 0.2170 -0.0338 0.8287
z.svi        0.7662 0.2471 0.2443  0.2806 1.2517
z.lcp        -0.4525 0.3950 0.3905 -1.2285 0.3235
z.gleason    0.1354 0.4779 0.4724 -0.8034 1.0742
z.pgg45      0.4525 0.4472 0.4421 -0.4261 1.3311

> ## Again, Bayes CI's are same (numerically) save as freq CIs:
> confint(zprostate.lm)

              2.5 %      97.5 %
(Intercept) -0.16636620 1.00923821
z.lcavol     2.13079625 3.93676294
z.lweight     0.43521996 2.95748666
z.age        -1.58994610 0.09751831
z.lbph       -0.03378987 0.82868898
z.svi        0.28064211 1.25166966
z.lcp        -1.22854622 0.32348608
z.gleason    -0.80337472 1.07419051
z.pgg45      -0.42608519 1.33114992

```

The following figure reproduces [Wak13, Fig. 5.10] (up to typos in his code and manifested in his figure, at least in my printing of his textbook). I offer my own code here, based on your textbook author's code, but you can see your textbook author's Web site for his original code, if you like. (His typos standardize all  $k = 8$  marginal  $t$  posteriors by the same posterior  $t$  scale hyperparameter for  $\beta_1$  (i.e., by  $\widehat{SE}(\widehat{\beta}_1)$ ). I correct the typo in my version of the code. (When I don't, I get his incorrect plot, of course; not shown here.)

To understand how the plots are created in the code, we should know that  $t$  distributions are what is called a **location-scale family** of distributions. For univariate  $t$  distributions, like our posterior marginals for the  $\beta_j$ , we have

$$t \sim t_1(\mu, \sigma^2, \nu),$$

where  $\mu$  is the **location** parameter, which, in the current context, corresponds to our posterior mean, i.e.,  $\mu = \widehat{\beta}_j$ , (when degrees of freedom  $\nu > 1$ ). And,  $\sigma$  corresponds to the **scale** parameter, which, in our cur-

rent context, corresponds to the posterior scale  $\sigma = \sqrt{\widehat{SE}(\widehat{\beta}_j)}$ , and note that  $\text{Var}(t) = \nu/(\nu - 2)\sigma^2$  (the scale is not the standard deviation).

If we denote  $f(t)$  to be the pdf of  $t$ , then

$$f(t) = \frac{1}{\sigma} h\left(\frac{t - \mu}{\sigma}\right)$$

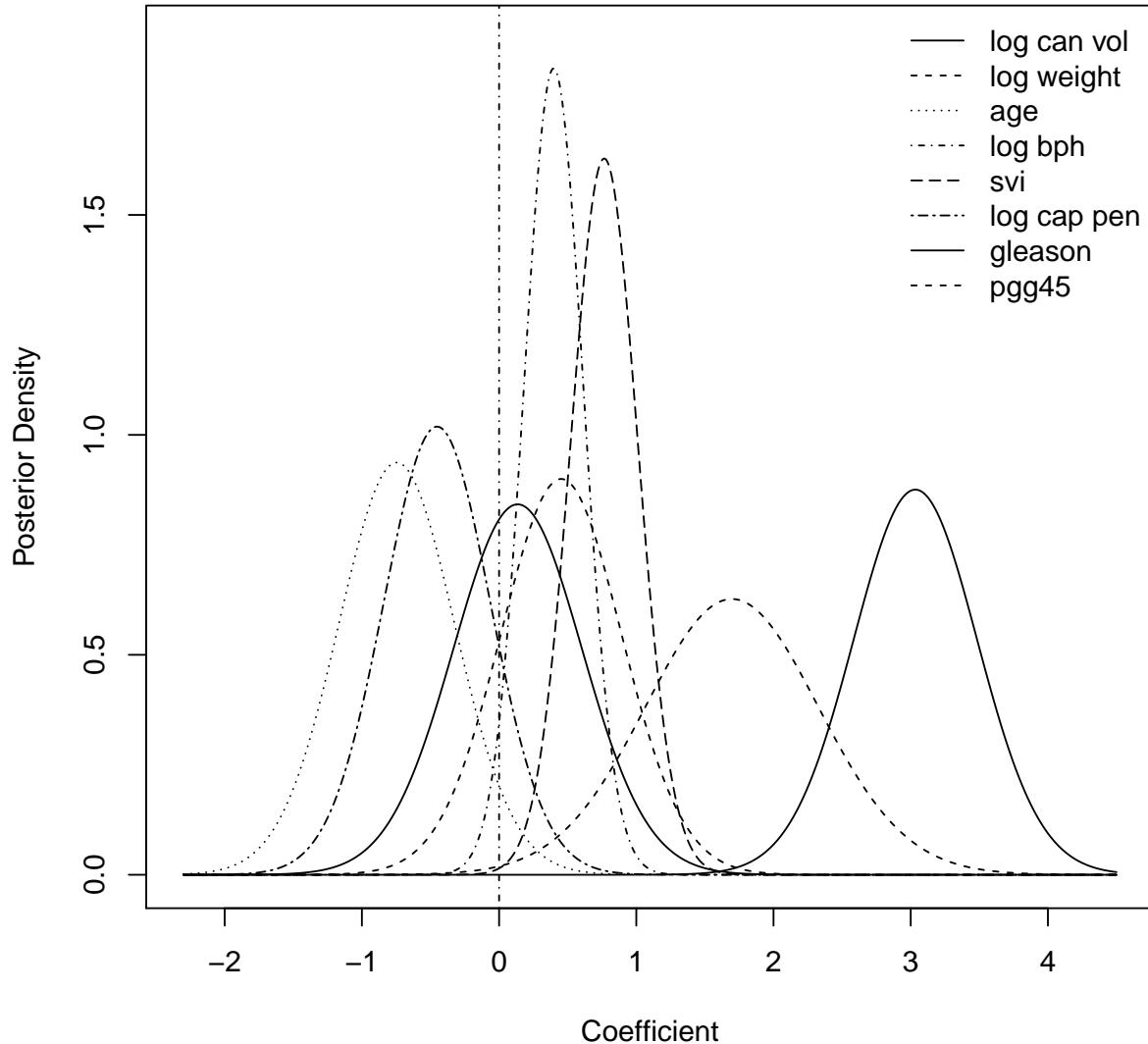
where

$$h(t) = t(0, 1, \nu)$$

is the pdf of Student's  $t$ , i.e., the standard  $t$  (analogous to  $N(0, 1)$ ). In other words, just as we can use a standard normal (table) to compute probabilities/quantiles (and densities) of general normal random variables, we can use a standard (Student's)  $t$  for more general  $t$  random variables from their location-scale family.

```
> ## Fig 5.10 -- posterior distributions under a flat prior
> ##
> par(mfrow=c(1,1),mar=c(5, 4, 4, 2) +0.1)
> ## Grid of beta values to plot marg. post. t densities
> bgrid<- seq(-2.3,4.5,.01)
> ## Omit intercept marginal (param=1) as in Fig. 5.10:
> invisible(sapply(2:9, FUN=function(param,bgrid,pmean,pscale,pdf){
+   ## Using location-scale properties of t distribution to plot:
+   tmargdens<- dt((bgrid - pmean[param])/pscale[param], df=pdf)/
+     pscale[param]
+   if(param==2)
+     plot(bgrid,tmargdens,type="l",xlab="Coefficient",
+           ylab="Posterior Density",
+           ylim=c(0,1.9), lty=param-1)
+   else
+     lines(bgrid,tmargdens,lty=param-1)
+ },
+ bgrid=bgrid,
+ pmean=tpostmeans,
+ pscale=sqrt(tpostscals2),
+ pdf=tpostpdf
+ ))
> abline(v=0,lty=4) ## not sure why 4
> legend("topright",legend=c("log can vol","log weight","age","log bph",
```

```
+ "svi", "log cap pen", "gleason", "pgg45"),
+ bty="n", lty=1:8)
```



[Wak13, Fig. 5.11] shows the 95% credible intervals graphically (solid lines) associated with the marginal pdfs in [Wak13, Fig. 5.10] (Evidently, his typos, mentioned above, do not affect the intervals (solid lines) in [Wak13, Fig. 5.11].) We will re-create [Wak13, Fig. 5.11] later, after discussing Gibbs sampling in the context of our next prior/posterior. Incidentally, he ([Wak13, Fig. 5.11]) uses integrated nested lapace approximations (INLA; [Wak13,

§3.7.4]) to the posteriors, an alternative way to deal with the integration problem related to the normalizing constant in Bayes theorem and hence to obtain the posterior intervals for the particular improper prior analysis, here, and for the next analysis using a different prior/posterior. We, of course, used the exact posterior here—not sure if INLA returns the exact credible intervals if they’re available. (This is why aforementioned code typos do not affect [Wak13, Fig. 5.11] as they did his [Wak13, Fig. 5.10].) And, for our next prior/posterior, we will use Gibbs sampling instead of INLA. In any case, again, we wait to re-create [Wak13, Fig.5.11].

## 8.9 A Common Independence Prior

$$\begin{aligned} [\boldsymbol{\beta}, \sigma^2] &= [\boldsymbol{\beta}][\sigma^2] \\ &= N(\mathbf{m}_0, \Sigma_0) \times \text{inv-}\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

- Notice, in particular, that we now assume **independence** a priori. See middle of [Wak13, p. 223]. (Again, an inverse gamma is equivalent to an  $\text{inv-}\chi^2$ , and saying  $\sigma^{-2} \sim \text{gamma}(\alpha, \beta)$  is the same as saying  $\sigma^2 \sim \text{inv-gamma}(\alpha, \beta)$ .)
- This now **does not give a closed form posterior**, like the two priors that we considered, above, unless we consider let  $\Sigma_0^{-1} = \mathbf{0}$  and  $\nu_0 = 0$ , which gives the previous improper prior and closed-form  $N \times \text{inv-}\chi^2$  posterior results.
- It’s **not a conjugate prior** (unless we consider the improper case as member of the normal-inv- $\chi^2$  family...).
- In other words, we do not get a recognizable posterior distribution (for which we usually know means, variances, probabilities, R functions, etc.).
- Now what? (wait a moment)

### 8.9.1 Full Conditional Posterior Distributions

We do however know the form of **full conditional posterior distributions** (often shortened to “**full conditionals**”)

$$\begin{aligned}\boldsymbol{\beta} | \sigma^2, \mathbf{y} &\sim N(\widehat{\mathbf{m}}, \sigma^2 \widehat{\Sigma}) \\ \sigma^2 | \boldsymbol{\beta}, \mathbf{y} &\sim \text{inv-}\chi^2(\widehat{\nu}, \widehat{\sigma}^2),\end{aligned}$$

where

$$\begin{aligned}\widehat{\mathbf{m}} &= (\mathbf{I} - \mathbf{W})\mathbf{m}_0 + \mathbf{W}\widehat{\boldsymbol{\beta}}, \\ \widehat{\Sigma} &= \mathbf{W}(\mathbf{X}^t \mathbf{X})^{-1} \\ \widehat{\nu} &= \nu_0 + n \\ \widehat{\sigma}^2 &= (1/\widehat{\nu})(\nu_0 \sigma_0^2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) \\ \mathbf{W} &= (\sigma^2 \Sigma_0^{-1} + \mathbf{X}^t \mathbf{X})^{-1}(\mathbf{X}^t \mathbf{X})\end{aligned}$$

- See [Wak13, Expr. (5.46) & (5.47)] for these full conditionals with a change of notation and use of inverse gamma instead of our inv- $\chi^2$ .
- Very often, we see the **particular prior** with  $\mathbf{m}_0 = \mathbf{0}$  and  $\Sigma_0 = \sigma_\beta^2 \mathbf{I}$ , i.e.,

$$[\boldsymbol{\beta}] = N(\mathbf{0}, \sigma_\beta^2 \mathbf{I}),$$

with the same inv- $\chi^2$  prior as just given above.

- In this **particular case**, we have the **full conditional**,

$$[\boldsymbol{\beta} | \sigma^2, \mathbf{y}] = N(\widehat{\mathbf{m}}, \sigma^2 \widehat{\Sigma})$$

where, now,

$$\begin{aligned}\widehat{\mathbf{m}} &= \widehat{\Sigma} \mathbf{X}^t \mathbf{y}, \\ \widehat{\Sigma} &= ((\sigma^2 / \sigma_\beta^2) \mathbf{I} + \mathbf{X}^t \mathbf{X})^{-1},\end{aligned}$$

and the other full conditional posterior remains the same as before.

- This gets us very close to typical (Bayesian) **shrinkage methods** and related methods, e.g., LASSO and ridge regression ([HTF01, §3.4], [Wak13, §10.5 & 10.6], [JWHT14, §6.2]), as well as to typical **Bayesian variable selection** methods (ref?).
- We never did answer the question, ‘Now what?’ That is, how does knowing the **full conditionals** get us closer to the posterior (and posterior predictive)? There are many answers to those questions, the most popular answer being **Gibbs sampling** or, more generally, Markov chain Monte Carlo (MCMC) methods (McMC?...). See [Wak13, §3.8].

## 8.10 2-Stage Gibbs Sampling

Given all full conditional distributions in a generic setting, your textbook’s author gives a Gibbs sampling algorithm in [Wak13, §3.8.4]. We restate it here for the two full conditionals obtained in the previous section with the latest prior under discussion.

For an initially chose value of  $\sigma^2 = \sigma^{2(0)}$ ,

1. sample  $\boldsymbol{\beta}^{(t+1)} | \sigma^{2(t)}, \mathbf{y} \sim [\boldsymbol{\beta} | \sigma^{2(t)}] = N(\widehat{\mathbf{m}}, \sigma^{2(t)} \widehat{\Sigma})$
2. sample  $\sigma^{2(t+1)} | \boldsymbol{\beta}^{(t+1)}, \mathbf{y} \sim [\sigma^2 | \boldsymbol{\beta}^{(t+1)}] = \text{inv-}\chi^2(\widehat{\nu}, \widehat{\sigma}^2)$
3. repeat 1 & 2 “to convergence” (to be discussed)

Note, we don’t need an initial value,  $\boldsymbol{\beta}^{(0)}$ , and, recall,  $\boldsymbol{\beta}$  is inside of  $\widehat{\sigma}^2$ .

## 8.11 Example

We continue our analysis of the prostate data, now in the context of the latest prior, full conditionals, and Gibbs sampling.

### 8.11.1 Eliciting a Prior

We need a specific form of prior in order to code specific full-conditional distributions for Gibbs sampling. We follow [Wak13, pp. 246-7] to construct a prior, which is a sort of combination of the independence prior and previous improper priors that we have discussed. Omitting details, the prior leads to known full-conditionals of the same form as above, but not to a known form for the posterior.

In particular, we (he) specify our prior as

$$[\boldsymbol{\beta}, \sigma^2] = [\boldsymbol{\beta}][\sigma^2] \propto \left( \prod_{j=0}^k [\beta_j] \right) \sigma^{-2},$$

where  $[\beta_0] \propto 1$ , improper for the intercept and for  $\sigma^2$ , and we (he) use informative priors,  $[\beta_j] = N(0, V)$ ,  $j > 0$ , where the prior standard deviation,  $\sqrt{V}$ , is chosen based on our (his) belief that it is unlikely that any of the standardized covariates, over their range of  $(0,1)$ , will change the median PSA by more than 10 units—equivalently, unlikely to change the mean  $\log(\text{PSA})$  by more than  $\log(10)$ . In other words, as a covariate ranges over  $(0,1)$ , we expect *a priori* that it is unlikely that  $|\beta_j|$  will exceed a change of  $\log(10)$ . To incorporate this into the normal prior with the *a priori* expectation of zero effect (on log scale), we (he) set(s)  $\log(10) = 1.96\sqrt{V}$ , and solves for  $V = (\log(10)/1.96)^2$ . (Why?) (Incidentally, in our (his INLA) code, we (he) use(s) the prior **precision**,  $1/V$ .)

**NOTE:** I follow the form of the full posterior distributions given in §8.9.1. I did not check, for the particular prior here, which is a sort of combination of previously discussed prior forms, that the full conditionals of §8.9.1 still hold, but I think they do. We'd have to do some derivations to verify this. In any case, we will see that our Gibbs results look very comparable the other analyses that we have done, above, and will do, below. In particular, our version of [Wak13, Fig 5.11] looks very comparable to that figure, where we see shrinkage towards the prior mean of zero for the informative (combo) prior considered here. We wait to produce our version of the figure until the end of this lecture chapter.

NOTE: I do not claim that the following code is somehow computationally efficient. But, its the sort of “hand” computations that, hopefully, illustrate our typeset material.

```

> y<- zprostate.lm$model[,1]
> X<- model.matrix(zprostate.lm)
> (n<- dim(X)[1])

[1] 97

> (p<- dim(X)[2])

[1] 9

> XtX<- t(X)%*%X
> bhat<- coef(zprostate.lm)
> nu0<- 0; ## prior df
> sig02<- 0; ## prior squared scale
> m0<- rep(0,p) ## prior beta mean
>
> ## prior precision (inverse variance) for beta
> V<- (log(10)/1.96)^2
> Sigma0inv<- diag(c(0,rep(1/V,p-1)))
>
> ## full cond. post. df for sigma2
> nuhat<- nu0 + n
>
> M<- 20000 ## MCMC iterations
> sigma2 <- rep(NA,M+1)
> beta<- matrix(NA,nrow=p,ncol=M+1)
> sigma2[1]<- summary(zprostate.lm)$sigma^2 ## initial sigma20
>
> library(mvtnorm)
> set.seed(20500 + 5150 + 86)
> for (i in 1:M){
+   ## intermediate ``weight'' matrix:
+   W<- solve(sigma2[i] * Sigma0inv + XtX)%*%XtX
+   ## full cond. post. mean for beta:
+   mhat<- (diag(p) - W)%*%m0 + W%*%bhat
+
+   ## full cond. post variance for beta up to sigma2:
+   Sighat<- W%*%solve(XtX)

```

```

+ ## generate beta / sigma2, y:
+ beta[,i+1]<- as.vector(rmvnorm(n=1,mean=mhat,sigma=sigma2[i]*Sighat))

+ ## generate sigma2 / beta, y
+ ## see BDA 3 appendix A for generating scaled inv-chi2 (see also
+ ## Wakefield's expression (5.45)...typo?):
+ sigma2hat<- (nu0*sig02 + sum((y - X%*%beta[,i+1])^2))/nuhat
+ sigma2[i+1]<- nuhat * sigma2hat / rchisq(n=1,df=nuhat)
+ }
> detach(package:mvtnorm)
> rm(y,p,n,X,XtX,bhat,nu0,sig02,m0,V,Sigma0inv,nuhat)

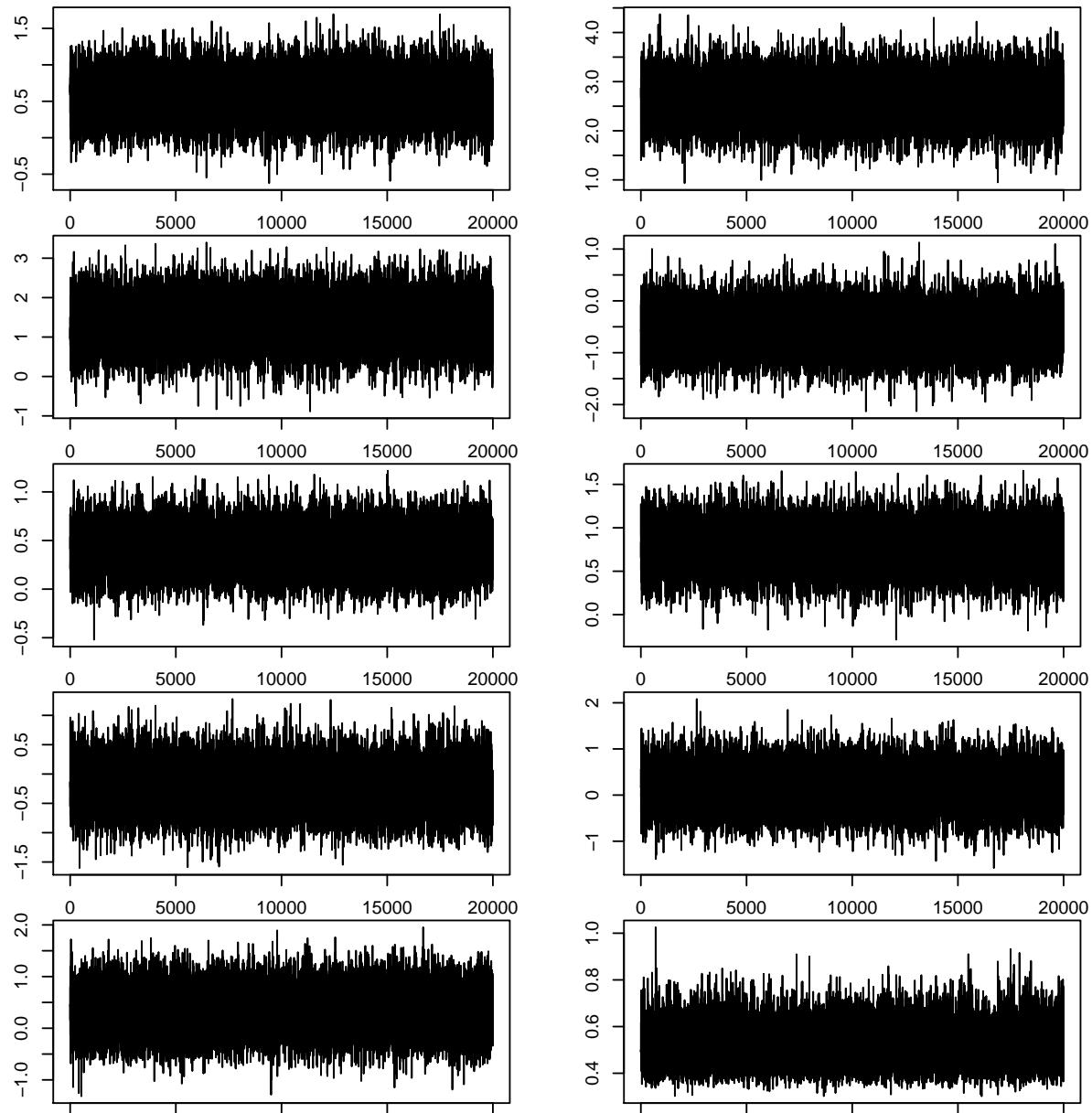
```

Some **history plots** to explore convergence. Ideally, we should have more than one chain. We obtain multiple chains in our Stan implementation, below.

```

> par(mfrow=c(5,2), plt=c(0.1,0.9,0.1,0.9))
> for(j in 1:9)
+   plot(0:M, beta[j,], type="l", xlab="MCMC iteration i",
+         ylab=expression(paste(beta[j], " | y")))
> plot(0:M, sigma2, type="l", xlab="MCMC iteration i",
+       ylab=expression(paste(sigma^2, " | y")))

```



- How do we know Gibbs sampling works? How do we know that the full conditionals are sufficient for getting samples from the (unknown) posterior?

## 8.12 Hamiltonian Monte Carlo in Stan

Above, we implemented Gibbs sampling ([Wak13, §3.8.4]) “by hand.” Gibbs sampling is special case of the Metropolis-Hastings algorithm ([Wak13, §3.8.2]). Both fall under the guise of Markov chain Monte Carlo sampling methods (MCMC; [Wak13, §3.8]). Here, we use another type of MCMC sampling known as Hamiltonian Monte Carlo (HMC) (sometimes “hybrid Monte Carlo,” but not in the same sense as [Wak13, §3.8.5]). We do not discuss the details of HMC, but merely use Stan, via the RStan interface, to implement HMC for the current running prior as we have been following in [Wak13, §5.12]. Of course, with large enough MCMC sample size, our Gibbs sampling results should be practically the same as the HMC results; we’re sampling from the same posterior, after all. But, you may find the relatively high level interface of Stan to be more convenient compared to programming the HMC algorithm or other MCMC yourself.

I do not discuss the preliminary details of installing Stan or the `rstan` package. You should be able to do that. The `rstan` package comes with its own documentation, as most any other R package. Also, the Stan language comes with its own manual, `stan-reference-2.17.0.pdf`, available on the Web. I use this manual a lot. Incidentally, there are other interfaces to the Stan language besides RStan, e.g., PyStan, for Python enthusiasts.

In the following subsections, we implement the current running Bayesian model example in Stan’s **program blocks**. Together, these blocks constitute a Stan language program, which will be parsed and translated to C++, then compiled and linked into object code, ready to be loaded and run. More discussion in class.

### 8.12.1 Functions Block

```
functions {
    // nothing for this example
}
```

### 8.12.2 Data Block

```
data{  
    int<lower=1> N;  
    int<lower=1> p;  
    matrix[N,p] X;  
    vector[N] y;  
    vector[p] m0;  
    matrix[p,p] prec0;  
    real<lower=0> nu0;  
    real<lower=0> sig20;  
}
```

### 8.12.3 Transformed Data Block

```
transformed data{  
    // nothing for this example  
}
```

### 8.12.4 Parameters Block

```
parameters{  
    vector[p] beta;  
    real lnsigma2;  
}
```

### 8.12.5 Transformed Parameters Block

```
transformed parameters{  
    real<lower=0> sigma2 = exp(lnsigma2);  
}
```

### 8.12.6 Model Block

Here, we are merely specifying the joint model in the numerator of Bayes theorem.

```
model{
    y ~ normal(X*beta, sigma2);
    // Excluding intercept beta0:
    segment(beta, 2, p) ~ multi_normal_prec(segment(m0, 2, p), block(prec0, 2, 2, p, p));

    *****
    Other parameters (in parameters block), without an explicit prior
    specification, here, will receive a uniform prior over their
    implied support. For us, this means [beta0] will be proportional
    to 1 over (-inf, inf), improper as desired, and that
    [lnsigma2] is proportional to 1 on (-inf, inf) (improper), which
    corresponds to [sigma2] proportional to 1/sigma2, improper as
    desired.

    *****/
}
```

### 8.12.7 Generated Quantities Block

```
generated quantities{
    // nothing for this example
}
```

### 8.12.8 Altogether for Stan

I have put all of the above program block code into one ASCII (text) file called `zprostate1.stan`. The next chunk reads and displays the file contents, with all of the program blocks together in a working Stan program.

```
> writeLines(readLines("./Stan/zprostate1.stan"))

functions {
  // nothing for this example
}

data{
  int<lower=1> N;
  int<lower=1> p;
  matrix[N,p] X;
  vector[N] y;
  vector[p] m0;
  matrix[p,p] prec0;
  real<lower=0> nu0;
  real<lower=0> sig20;
}

transformed data{
  // nothing for this example
}

parameters{
  vector[p] beta;
  real lnsigma2;
}

transformed parameters{
  real<lower=0> sigma2 = exp(lnsigma2);
}

model{
  y ~ normal(X*beta, sigma2);
  // Excluding intercept beta0:
  segment(beta,2,p-1) ~ multi_normal_prec(segment(m0,2,p-1), block(prec0,2,2,p-1,p-1));

  *****
}
```

Other parameters (in parameters block), without an explicit prior specification, here, will receive a uniform prior over their implied support. For us, this means [beta0] will be proportional to 1 over (-inf,inf), improper as desired, and that that m[lnsigma2] is proportional to 1 on (-inf, inf) (improper), which corresponds to [sigma2] proportional to 1/sigma2, improper as

```

desired.

*****
}

generated quantities{
  // nothing for this example
}

```

### 8.12.9 Translate Stan to C++ with stanc

The next chunk uses the `rstan` function `stanc` to parse and translate Stan code in `zprostate1.stan` into a C++ file, checking for Stan syntax errors. Address any errors reported by editing the Stan code in `zprostate1.stan` until `stanc` returns no errors/warnings.

```

> library(rstan,quietly=TRUE)

rstan (Version 2.17.3, GitRev: 2e1f913d3ca3)
For execution on a local, multicore CPU with excess RAM we recommend calling
options(mc.cores = parallel::detectCores()).
To avoid recompilation of unchanged Stan programs, we recommend calling
rstan_options(auto_write = TRUE)

> options(mc.cores = parallel::detectCores())
> rstan_options(auto_write = TRUE)
> zprostate1.stanc<- stanc(file="./Stan/zprostate1.stan")

```

### 8.12.10 Make an Executable Stan Model with stan\_model

The next chunk compiles the C++ code in the object returned by `stanc`, above, into a `stanmodel` object, which references compiled/linked object code ready to be (re)used in sampling, below, with the `sampling` function in `rstan`.

```
> zprostate1.stanmod<- stan_model(stanc_ret=zprostate1.stanc)
```

### 8.12.11 Data List for Stan

Now we are ready to create data and initial value lists to pass to our Stan model to obtain samples from the posterior distribution. If we do not specify initial values, then Stan generates these, which may work fine. In the case when Stan's initial values cause sampling to fail, then you should pass (better) initial values to Stan in a list.

We use a previous `lm` object to help us along.

```
> X<- model.matrix(zprostate.lm)
> zprostate1.data<- list(
+   N=dim(X)[1],
+   p=dim(X)[2],
+   X=X,
+   y=zprostate.lm$model[,1],
+   m0=rep(0, dim(X)[2]),
+   prec0=diag(c(0,rep((log(10)/1.96)^2, dim(X)[2]-1))),
+   nu0=0.0,
+   sig20=0.0)
> rm(X)
```

### 8.12.12 List of Initial Value Lists for Stan

```
> library(mvtnorm)
> set.seed(20500 + 5150 + 24601)
> binit<- rmvnorm(n=3,mean=coef(zprostate.lm),sigma=3*vcov(zprostate.lm))
> attach(zprostate1.data)
```

*The following object is masked by .GlobalEnv:*

*N*

```
> sigma2init<- NULL
> sigma2init<- c(sigma2init, sum((y - X%*%binit[1,])^2)/(N-p))
> sigma2init<- c(sigma2init, sum((y - X%*%binit[2,])^2)/(N-p))
> sigma2init<- c(sigma2init, sum((y - X%*%binit[3,])^2)/(N-p))
> detach(zprostate1.data)
> zprostate1.init<- list(
+   list(beta=binit[1,], lnsigma2=log(sigma2init[1])),
+   list(beta=binit[2,], lnsigma2=log(sigma2init[2])),
+   list(beta=binit[3,], lnsigma2=log(sigma2init[3])))
```

```
> detach(package:mvtnorm)
> rm(binit, sigma2init)
```

### 8.12.13 Executing a Stan Model with sampling

```
> zprostate1.fit<- sampling(zprostate1.stanmod,
+                               data=zprostate1.data,
+                               pars=c("beta", "sigma2"),
+                               chains=3,
+                               iter=10000,
+                               warmup=5000,
+                               init=zprostate1.init,
+                               refresh=1000)
> ## Good idea to save the fit to a file for later use:
> save(list=c("zprostate1.fit"), file=".~/Stan/zprostate1.fit.RData")
```

### 8.12.14 Posterior Summaries with coda

```
> load(file=".~/Stan/zprostate1.fit.RData")
> ## Transform Stan model fit (stanfit class) to a coda mcmc.list object
> ## for use in coda:
> zprostate1.mcmc<- As.mcmc.list(zprostate1.fit)
> library(coda)
```

*Attaching package: 'coda'*  
*The following object is masked from 'package:rstan':*  
*traceplot*

```
> nchain(zprostate1.mcmc)
```

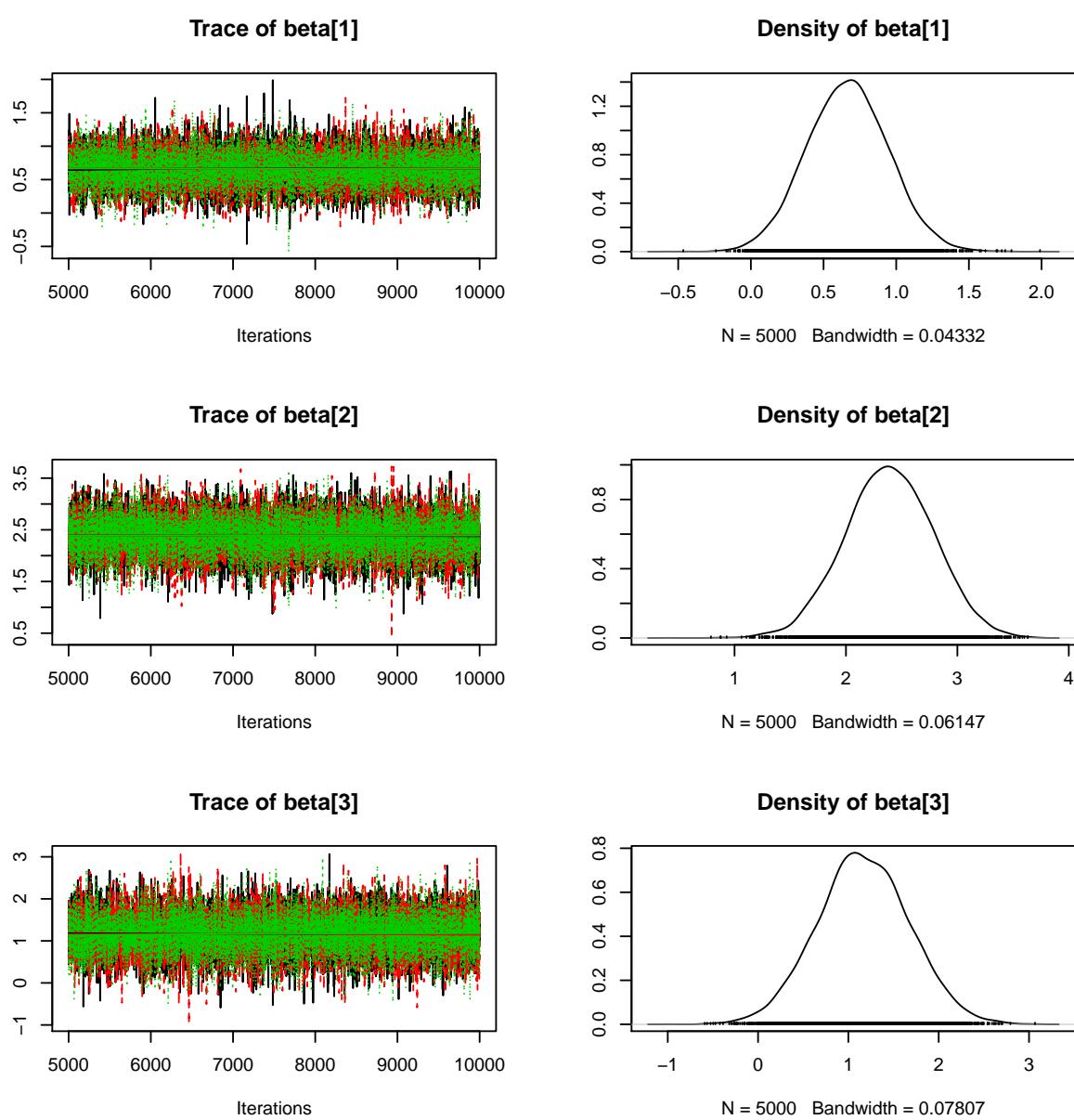
```
[1] 3
```

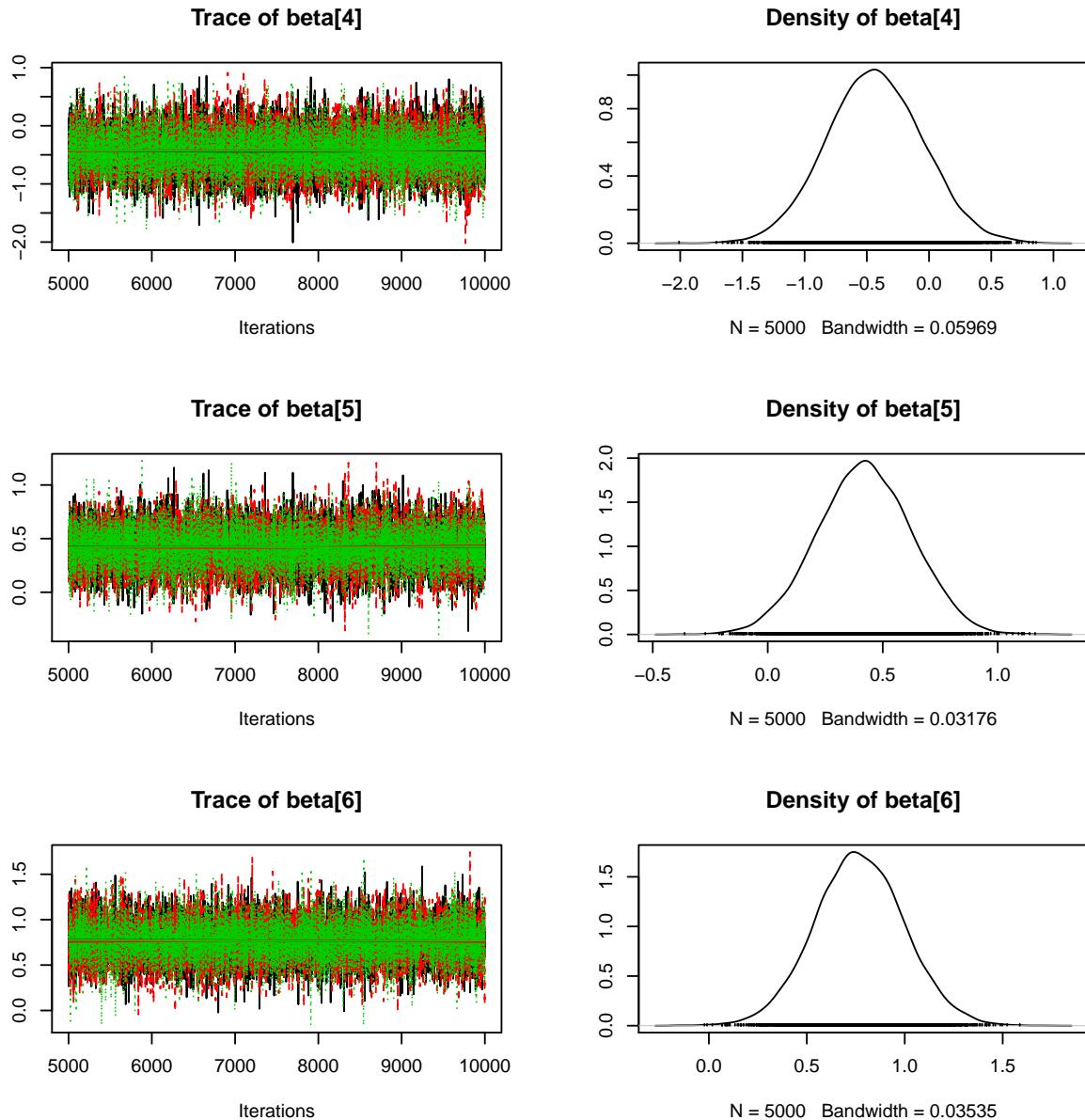
```
> niter(zprostate1.mcmc)
```

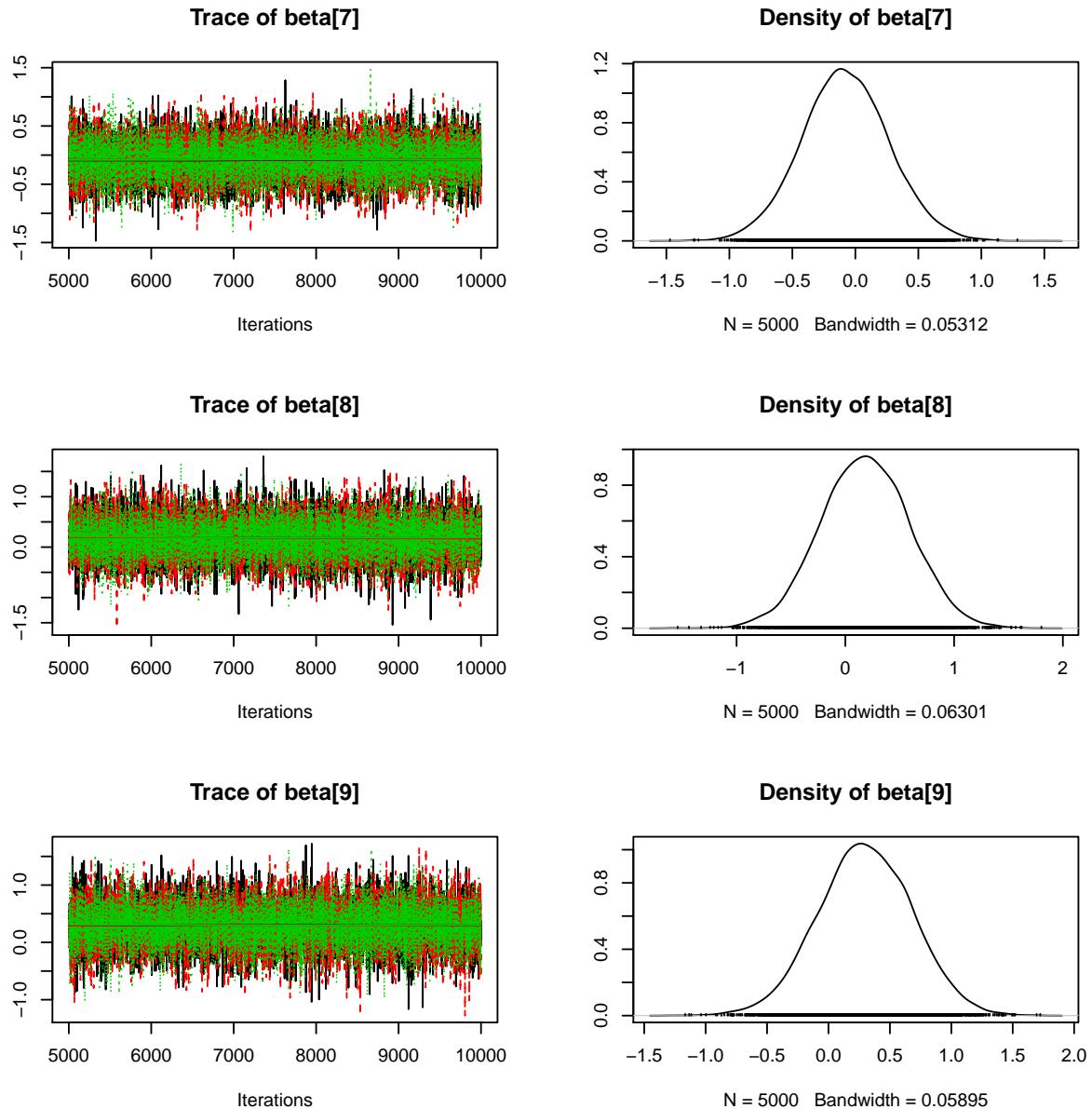
```
[1] 5000
```

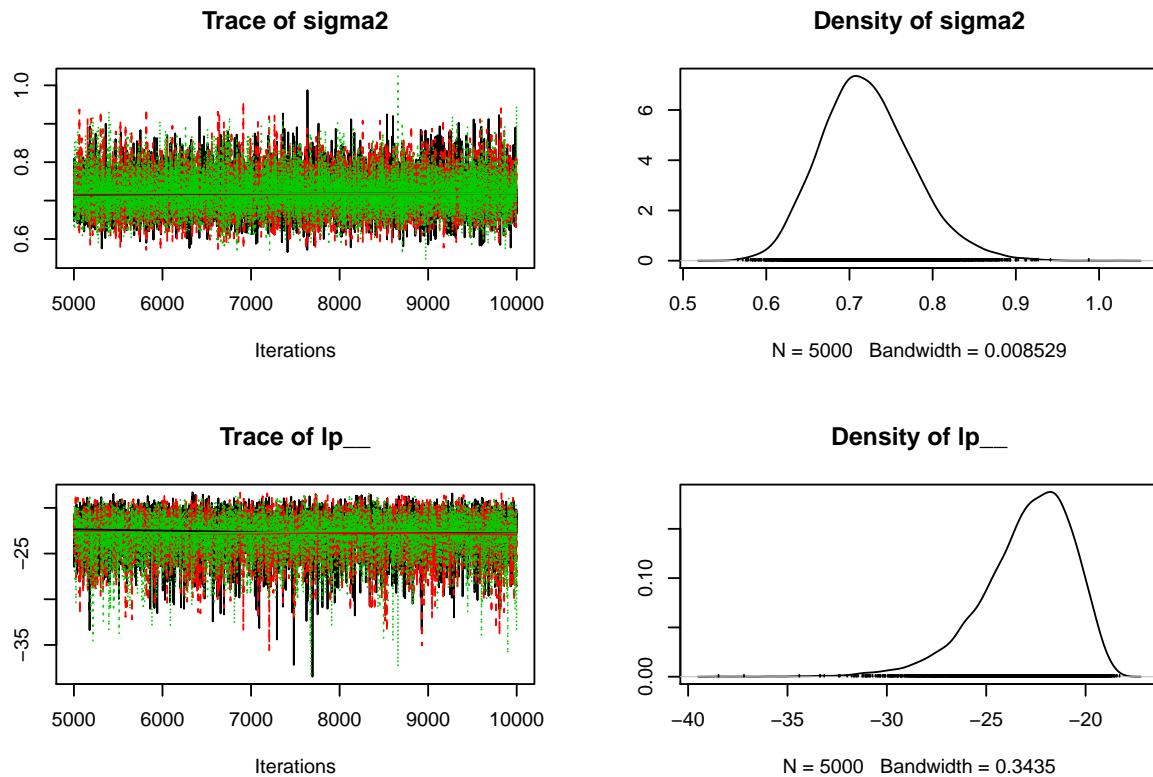
```
> nvar(zprostate1.mcmc)
```

```
[1] 11  
  
> varnames(zprostate1.mcmc)  
  
[1] "beta[1]" "beta[2]" "beta[3]" "beta[4]" "beta[5]"  
[6] "beta[6]" "beta[7]" "beta[8]" "beta[9]" "sigma2"  
[11] "lp_-"  
  
> plot(zprostate1.mcmc)
```









```
> (zprostate1.psum<- summary(zprostate1.mcmc))
```

```
Iterations = 5001:10000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 5000
```

1. Empirical mean and standard deviation for each variable,

plus standard error of the mean:

|         | Mean      | SD      | Naive SE  | Time-series SE |
|---------|-----------|---------|-----------|----------------|
| beta[1] | 0.66610   | 0.27964 | 0.0022832 | 0.0024514      |
| beta[2] | 2.38684   | 0.39679 | 0.0032398 | 0.0033212      |
| beta[3] | 1.16602   | 0.50865 | 0.0041531 | 0.0041095      |
| beta[4] | -0.43486  | 0.38534 | 0.0031463 | 0.0031348      |
| beta[5] | 0.42051   | 0.20504 | 0.0016741 | 0.0015768      |
| beta[6] | 0.76876   | 0.22908 | 0.0018705 | 0.0018013      |
| beta[7] | -0.07929  | 0.34461 | 0.0028137 | 0.0028556      |
| beta[8] | 0.18085   | 0.40677 | 0.0033212 | 0.0032621      |
| beta[9] | 0.29772   | 0.38057 | 0.0031073 | 0.0031489      |
| sigma2  | 0.72137   | 0.05574 | 0.0004551 | 0.0004483      |
| lp__    | -22.93773 | 2.32729 | 0.0190022 | 0.0299989      |

2. Quantiles for each variable:

|         | 2.5%      | 25%       | 50%       | 75%      | 97.5%    |
|---------|-----------|-----------|-----------|----------|----------|
| beta[1] | 0.12216   | 0.47630   | 0.66662   | 0.8542   | 1.2183   |
| beta[2] | 1.60908   | 2.12151   | 2.38712   | 2.6589   | 3.1557   |
| beta[3] | 0.18537   | 0.83071   | 1.15821   | 1.5060   | 2.1609   |
| beta[4] | -1.18441  | -0.69463  | -0.43778  | -0.1746  | 0.3268   |
| beta[5] | 0.01429   | 0.28316   | 0.42046   | 0.5587   | 0.8166   |
| beta[6] | 0.32140   | 0.61606   | 0.76629   | 0.9218   | 1.2237   |
| beta[7] | -0.75694  | -0.31031  | -0.08295  | 0.1492   | 0.6107   |
| beta[8] | -0.61526  | -0.09474  | 0.18031   | 0.4590   | 0.9751   |
| beta[9] | -0.45718  | 0.04482   | 0.29753   | 0.5593   | 1.0338   |
| sigma2  | 0.62309   | 0.68262   | 0.71787   | 0.7564   | 0.8422   |
| lp__    | -28.41158 | -24.23117 | -22.60886 | -21.2598 | -19.4536 |

## 8.13 Example Summary

Here, we compare the results from the improper prior analysis (§8.6 & 8.8) from Gibbs sampling (§8.10 & 8.11) and from HMC, just above, to get our version of [Wak13, Fig. 5.11], a display of Bayesian credible intervals for each analysis. Again, up to MCMC error, we expect the Gibbs results to be the same as the HMC results as both methods sample from the same posterior.

Note the **shrinkage** toward zero of the posteriors obtained from the informative prior (with mean zero).

```

> ## Use the previously created objects from our improper prior analysis:
> p1lo<- tpost25lb[2:9]
> p1hi<- tpost975ub[2:9]
>
> ## Omit the initial value and first 5000 samples, i.e., omit
> ## ``burn-in'' or ``warm-up'' iterations from Gibbs:
> p2lo<- apply(beta[2:9,-c(1:5001)], 1, quantile, probs=c(0.025))
> p2hi<- apply(beta[2:9, -c(1:5001)], 1, quantile, probs=c(0.975))
>
> ## From Stan HMC (already omitted first 5000 iterations from each of
> ## three chains):
> names(zprostate1.psum)

[1] "statistics" "quantiles" "start"      "end"
[5] "thin"        "nchain"

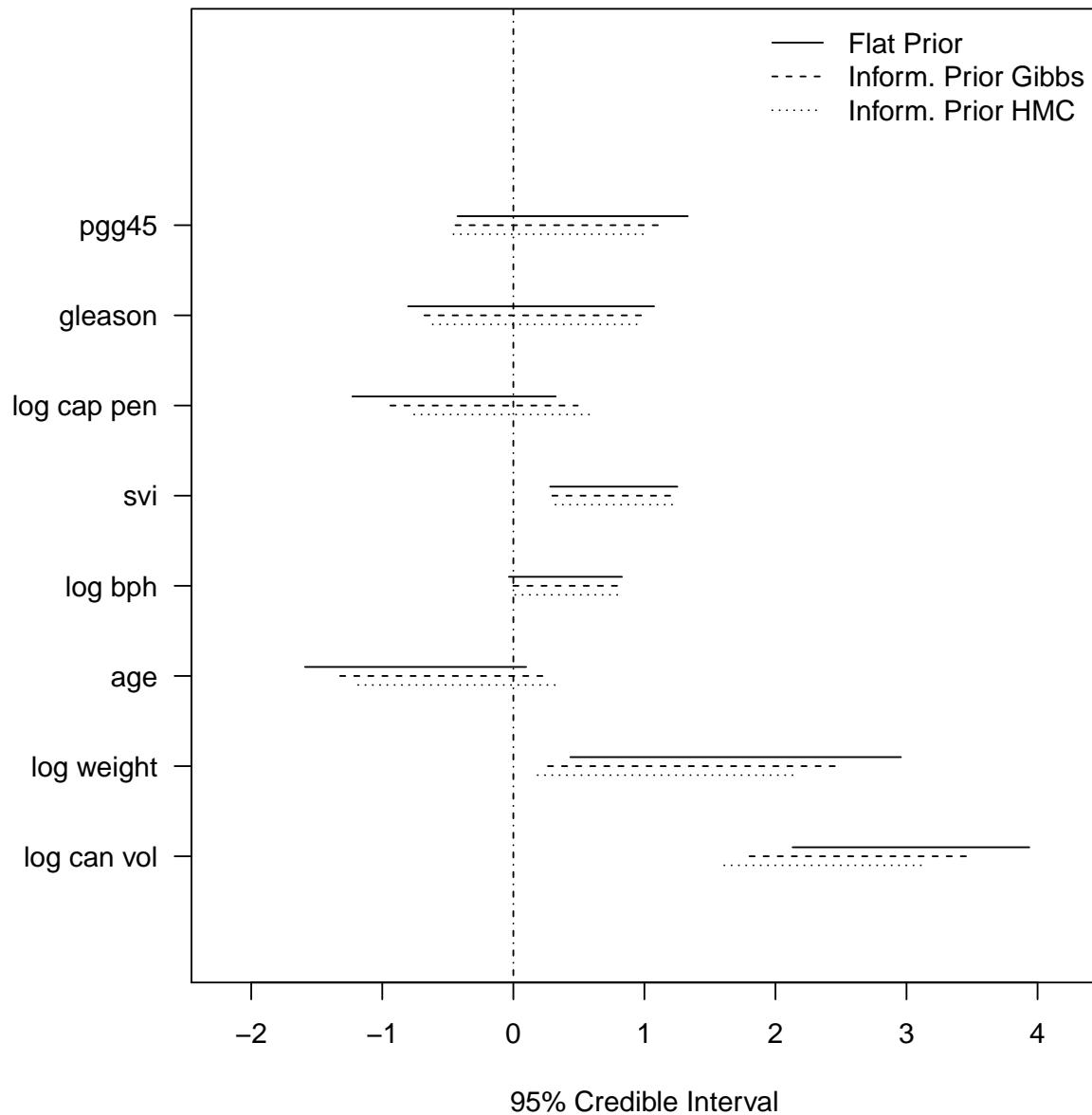
> p3lo<- zprostate1.psum$quantiles[2:9,c("2.5%")]
> p3hi<- zprostate1.psum$quantiles[2:9,c("97.5%")]

```

```

> ## Fig 5.11
> par(mar=c(5,7,1,1)+.1)
> plot(p1lo,p1hi,xlim=c(-2.2,4.2),type="n",xlab="95% Credible Interval",
+       ylab="",axes="F",ylim=c(0,10))
> box()
> axis(1)
> axis(2,at=seq(1,8),labels=c("log can vol","log weight","age","log bph","svi",
+                               "log cap pen","gleason","pgg45"),las=1)
> for (i in 1:8){
+   lines(y=c(i+.1,i+.1),x=c(p1lo[i],p1hi[i]))
+   lines(y=c(i,i),x=c(p2lo[i],p2hi[i]),lty=2)
+   lines(y=c(i-.1,i-.1),x=c(p3lo[i],p3hi[i]),lty=3)
+
+ }
> abline(v=0,lty=4)
> legend("topright",
+        legend=c("Flat Prior","Inform. Prior Gibbs", "Inform. Prior HMC"),
+        lty=1:3,bty="n")

```



```
> detach(package:coda)
> detach(package:rstan)
```

## 8.14 Other Priors

Generally speaking, for other priors on the parameters of the normal linear model, we will typically not recognize the posterior or full conditionals, in

which case we have to appeal to other methods, MCMC again being very popular, e.g., (Metropolis-)Hastings. See [Wak13, §3.8].

# Lecture 9

## One-Way ANOVA

### Contents

---

|        |  |     |
|--------|--|-----|
| 9.1    | Initial Concepts and Notation . . . . .                          | 331 |
| 9.2    | Cell Means (Regression) Model: $E(Y_{ij}) = \mu_i$ . . . . .     | 333 |
| 9.2.1  | Example . . . . .  | 337 |
| 9.3    | Cell Means Model: Further Inference About Means . . . . .        | 344 |
| 9.3.1  | Example . . . . .  | 345 |
| 9.4    | Factor Effects Parameterization . . . . .                        | 347 |
| 9.4.1  | Defining a Factor Effects Parameterization Using Treatment Means | 348 |
| 9.5    | Factor Effects Parameterization: Before Constraints . . . . .    | 349 |
| 9.6    | Imposing Constraints . . . . .                                   | 350 |
| 9.7    | Sum-to-Zero Constraint/Coding . . . . .                          | 351 |
| 9.8    | Reference Treatment Constraint/Coding . . . . .                  | 360 |
| 9.9    | Further Inference About Treatment Means . . . . .                | 368 |
| 9.9.1  | Example . . . . .  | 369 |
| 9.10   | Summary of One-Way ANOVA . . . . .                               | 373 |
| 9.10.1 | Regression Approach to ANOVA . . . . .                           | 376 |
|        | Model, Parametrization, Reparameterization . . . . .             | 376 |

---

*Main Objectives:*

- Cell means model for one-way ANOVA

- Factor effects model with sum-to-zero constraint/coding/parameterization and with treatment/cell reference/corner-point constraint/coding/parameterization
  - Learn how to use R to make inferences under different constraints/codings/parameterizations
- 
- $\mathcal{O}$

***Additional Reading:***

[Wak13, §5.8.1] (very brief!)

[RS13, Chap. 5 & 6]

[KNNL05, Chap. 16 & 17] \_\_\_\_\_  $\mathcal{R}$

Please, always be prepared to take additional notes in class!

## 9.1 Initial Concepts and Notation

First, we cover some useful notation/concepts for one-way ANOVA that will lead naturally to similar notation/concepts in for higher-way ANOVA, later. We've seen a bit of this before when discussing factor (qualitative) covariates in regression (§7.13). Of course, ANOVA is just a special case of our linear model, but this perspective seems overly simplistic given ANOVA's widespread use to simplify the understanding of relatively complicated data (not much in this class), particularly in the design of randomized experiments, which is far too much detail for us to cover. We only scratch the surface in some sense despite viewing ANOVA simply as a linear model, which hopefully helps us to understand ANOVA to some extent.

- **Number of factor variables:**  $[1]$ . Call this factor, generically, Factor A. We will consider multiple factor variables when we get to multi-way ANOVA, later, where this notation will extend generically and naturally: Factor A, Factor B, Factor C, etc. **What is a factor?**
- **Number of factor levels:**  $[a]$ . Again, when we get to multi-way ANOVA, this notation will extend to indicate number of levels for

each factor under consideration:  $a, b, c$ , etc. **What is a factor level?**

- **Treatments.** Treatments are the set of conditions defined by the unique combinations of factors levels across all factors. Here, we have only one factor, so that factor levels and treatment levels are synonymous. This will not generally be the case for higher-way ANOVA.
- **Sample sizes** (or number of units per treatment level) are denoted  $[n_i], i = 1, \dots, a$ .
- **Observation**  $[Y_{ij}]$  is the jth observation (response) in the ith treatment level,  $i = 1, \dots, a, j = 1, \dots, n_i$
- **Total number of observations**, i.e., total samples size, is denoted as  $[n_T]$ , i.e.,  $n_T = n_1 + \dots + n_a = \sum_{i=1}^a n_i$ .
- **Balance.** If we have an equal number of observations per treatment level, then we say our treatment design is balanced. Otherwise it is unbalanced. i.e.,  $n_1 = n_2 = \dots = n_a = n$ , where  $[n]$  is the common number of observations in each treatment. Thus, in the balanced case,  $[n_T = na]$ . (Note that we distinguish  $n$  from  $n_T$ .) Balance plays a relatively small role in the one-way case compared to the multi-way case, where it has historically receive much more attention.
- **Observational or experimental treatment level means** are denoted by  $[\mu_i], i = 1, \dots, a$ . These are unknown parameters to be estimated. These are not averages of the observations,  $Y_{ij}$ . Again, for one-way ANOVA, factor level means are synonymous with treatment level means. **What is an observational study? An experimental study?** (Recall note chapter 1.)
- **Scope of Inference.** (Recall note chapter 1.) Ideally, but often not in practice, we can define a “population” of interest from which we obtain a random sample. (We devote little attention to sampling.) In this case, it is generally accepted that (good) statistical inference

based on such a random sample from a well-defined population applies to the population. In other words, our scope of inference is more broadly applicable beyond just the units/subjects for which we have observations to the units/subjects of the entire population. Without such a random sample from well-defined population, (the scope of) our inferences may be restricted, perhaps to merely the group of units/subjects for which we have measurements! Relatedly, randomization of units/subjects to treatments (or vice-versa) is the gold standard of causal inference. In other words, it is generally accepted to make cause-effect statements about treatments causing observed responses in an experimental study. On the other hand, in an observational study, wherein no such randomization has occurred, preferred practice dictates that we do not claim causality, but, much less, only an association or correlation among treatments and responses.

## 9.2 Cell Means (Regression) Model: $E(Y_{ij}) = \mu_i$

Just as in regression, we build a **model** for the mean of our observations i.e., for  $E(Y_{ij})$ , i.e., for our **regression function**, now in the context of our specialized one-way ANOVA notation. The cell means model is particularly simple:

$$E(Y_{ij}) = \mu_i.$$

That is, the observations in the  $i$ th treatment (factor) level have their own mean, symbolized by a single parameter,  $\mu_i$ , for each level. Unlike regression, we do not constrain (inform?) the mean to fall on a (curvi)linear function of the covariate(s) (which is difficult to conceive for a categorical regressor (factor)!). Here, what serves to indicate the different values of the covariate?

More completely, we specify the cell means model by

$$Y_{ij} \stackrel{\text{ind}}{\sim} N(\mu_i, \sigma^2) \quad i = 1, \dots, a, \quad j = 1, \dots, n_i$$

or, equivalently (recall basic results in Lecture 3),

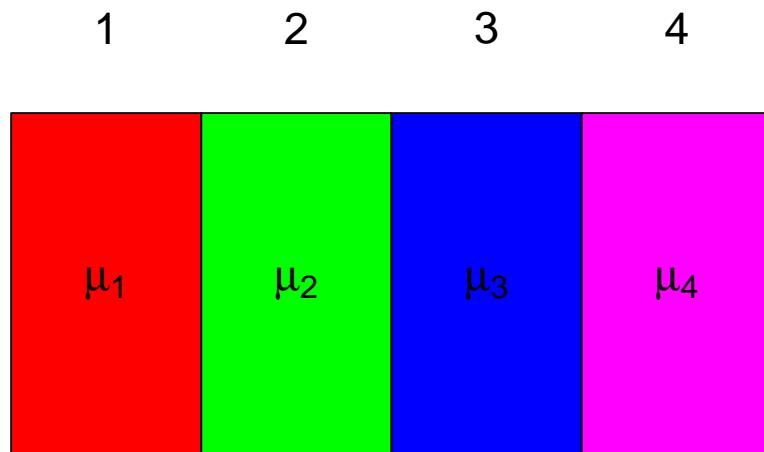
$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \epsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad i = 1, \dots, a, \quad j = 1, \dots, n_i$$

where  $\epsilon_{ij}$  is the **error** of observation  $j$  in treatment  $i$  and its (error) variance (component),  $\sigma^2$ , is assumed to be common to all treatment levels; in this latter sense, our procedures are “pooled” as in pooled  $t$  procedures that assume a common variance among treatment levels; we did not cover introductory statistics  $t$  procedures.

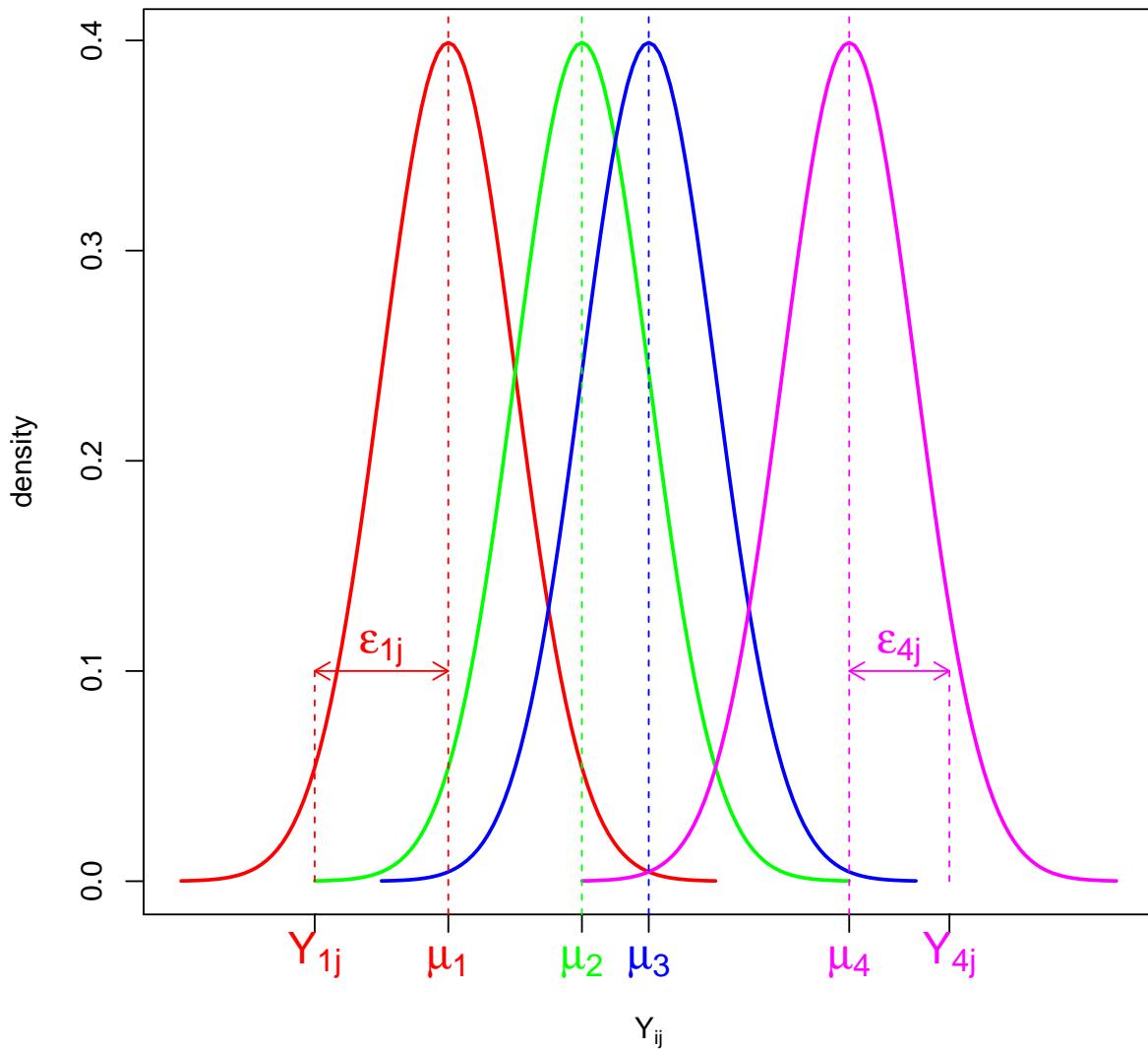
Also, note that the cell means model may be useful in higher-way ANOVA, where, in that case, cells (treatments) are defined by the combination of the levels of multiple factors, and notation may change slightly.

The one-way cell means layout is depicted in the next Chunk—far too simple, I know! Cells. Means. Oh my!

## Treatment levels



The next Chunk illustrates another way to envision the cell means model. (Note that my apparent ordering of means with factor (treatment) levels is purely accidental; it means nothing.)



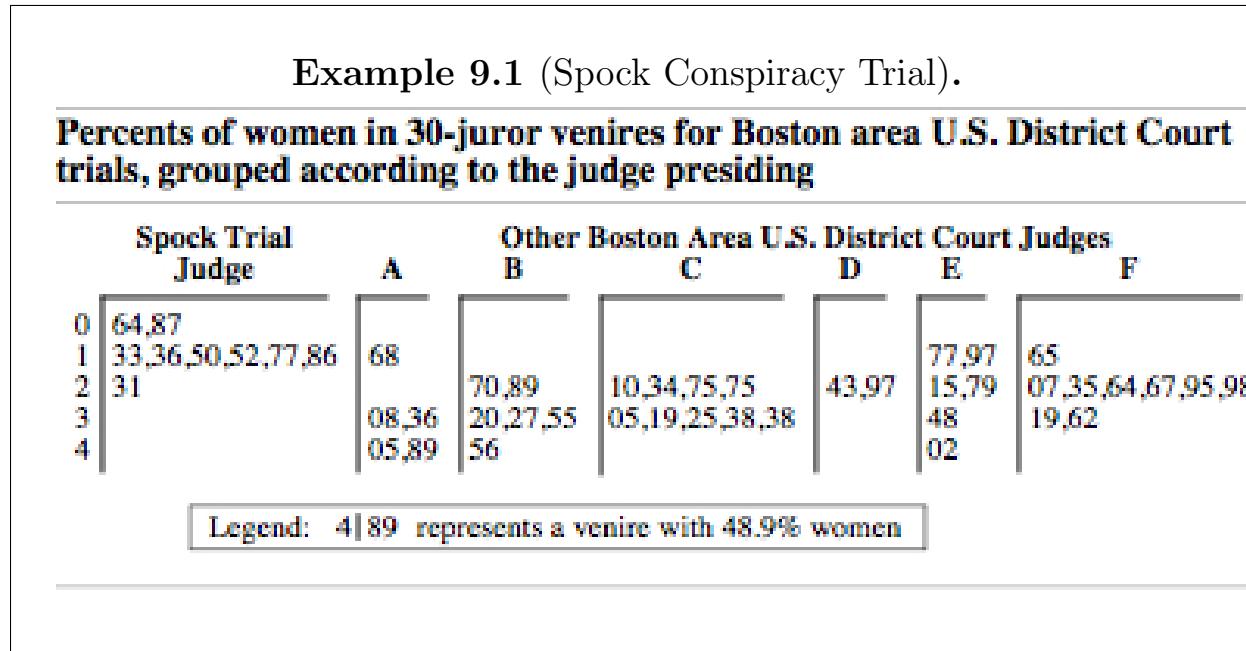
As we know, from Lecture 4, Example 4.22, we can write the cell means model in the form of a general linear model using matrix notation.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_T}).$$

We will write the elements of each of the vectors/matrices, above, on the board, in class. Please be prepared to take additional notes!

### 9.2.1 Example

The next example is taken from [RS13, Chap.5]. We will use it throughout our introduction to one-way ANOVA.



The next Chunk reads the example's data from Sleuth3 package and looks at it briefly. Note that R already views the Judge variable as a "Factor."

```
> case0502.df<- Sleuth3::case0502
> ##
> ## How is the data ``structured?:
> str(case0502.df)

'data.frame': 46 obs. of  2 variables:
 $ Percent: num  6.4 8.7 13.3 13.6 15 15.2 17.7 18.6 23.1 16.8 ...
 $ Judge   : Factor w/ 7 levels "A","B","C","D",...: 7 7 7 7 7 7 7 7 7 1 ...

> ## Summarize responses by Judge factor level (just lookin'):
> by(data=case0502.df, INDICES=case0502.df$Judge,
+     FUN=function(x) x$Percent)

case0502.df$Judge: A
[1] 16.8 30.8 33.6 40.5 48.9
```

```
case0502.df$Judge: B
[1] 27.0 28.9 32.0 32.7 35.5 45.6
-----
case0502.df$Judge: C
[1] 21.0 23.4 27.5 27.5 30.5 31.9 32.5 33.8 33.8
-----
case0502.df$Judge: D
[1] 24.3 29.7
-----
case0502.df$Judge: E
[1] 17.7 19.7 21.5 27.9 34.8 40.2
-----
case0502.df$Judge: F
[1] 16.5 20.7 23.5 26.4 26.7 29.5 29.8 31.9 36.2
-----
case0502.df$Judge: Spock's
[1] 6.4 8.7 13.3 13.6 15.0 15.2 17.7 18.6 23.1
```

Continuing our example, the next Chunk performs a one-way ANOVA in R using the cell means model defined above. It illustrates the use of an R model formula,

$$\text{Percent} \sim \text{Judge} - 1,$$

generically,

$$\text{Response} \sim \text{Expression},$$

just like we've seen for regression model formulas. By default, R will construct an  $\mathbf{X}$  matrix using a column of 1's, for an intercept parameter, as we've seen with regression. Our cell means model does not have an explicit intercept parameter, so we use “-1” to tell R not to code  $\mathbf{X}$  for an intercept parameter in  $\beta$ , i.e., to exclude the column of 1's from its  $\mathbf{X}$  matrix. (Alternatively, we could have specified  $\text{Percent} \sim \text{Judge} + 0$ .) With the 1's covariate excluded, R will attempt to continue to construct  $\mathbf{X}$  using a column of 0/1's for each level of each factor included in Expression, which includes only Judge in our current example. **What columns will R use for the Judge factor?**

**More in class.**

If there are no redundancies among the columns of the resulting  $\mathbf{X}$  matrix, then we have nothing further to consider for the  $\mathbf{X}$  matrix. Here, because we exclude the column of 1's, the  $\mathbf{X}$  matrix has no redundancy among its

columns (no linear dependencies; see note Section 4.2.6), and we (and R) need not consider further changes to  $\mathbf{X}$ : the resulting  $\mathbf{X}$  matrix is full rank, there is no redundancy amongst columns, and R can solve for the LS estimates of the  $\mu_i$  parameters as we discussed in §6.2. We will have more to say about handling redundancy in the  $\mathbf{X}$  matrix later.

```
> case0502cell.lm<- lm(Percent ~ Judge - 1, data=case0502.df)
>
> ## Let's look at the X matrix along side Judge:
> tmp<- cbind.data.frame(model.matrix(case0502cell.lm),
+                         Judge=case0502.df$Judge)
> names(tmp)<- c(LETTERS[1:6], "Spock's", "Judge")
> tmp

   A B C D E F Spock's     Judge
1  0 0 0 0 0 0      1 Spock's
2  0 0 0 0 0 0      1 Spock's
3  0 0 0 0 0 0      1 Spock's
4  0 0 0 0 0 0      1 Spock's
5  0 0 0 0 0 0      1 Spock's
6  0 0 0 0 0 0      1 Spock's
7  0 0 0 0 0 0      1 Spock's
8  0 0 0 0 0 0      1 Spock's
9  0 0 0 0 0 0      1 Spock's
10 1 0 0 0 0 0      0      A
11 1 0 0 0 0 0      0      A
12 1 0 0 0 0 0      0      A
13 1 0 0 0 0 0      0      A
14 1 0 0 0 0 0      0      A
15 0 1 0 0 0 0      0      B
16 0 1 0 0 0 0      0      B
17 0 1 0 0 0 0      0      B
18 0 1 0 0 0 0      0      B
19 0 1 0 0 0 0      0      B
20 0 1 0 0 0 0      0      B
21 0 0 1 0 0 0      0      C
22 0 0 1 0 0 0      0      C
23 0 0 1 0 0 0      0      C
24 0 0 1 0 0 0      0      C
25 0 0 1 0 0 0      0      C
26 0 0 1 0 0 0      0      C
27 0 0 1 0 0 0      0      C
```

```

28 0 0 1 0 0 0      0      C
29 0 0 1 0 0 0      0      C
30 0 0 0 1 0 0      0      D
31 0 0 0 1 0 0      0      D
32 0 0 0 0 1 0      0      E
33 0 0 0 0 1 0      0      E
34 0 0 0 0 1 0      0      E
35 0 0 0 0 1 0      0      E
36 0 0 0 0 1 0      0      E
37 0 0 0 0 1 0      0      E
38 0 0 0 0 0 1      0      F
39 0 0 0 0 0 1      0      F
40 0 0 0 0 0 1      0      F
41 0 0 0 0 0 1      0      F
42 0 0 0 0 0 1      0      F
43 0 0 0 0 0 1      0      F
44 0 0 0 0 0 1      0      F
45 0 0 0 0 0 1      0      F
46 0 0 0 0 0 1      0      F

```

```

> ## Estimated coefficients given by default:
> case0502cell.lm

```

Call:

```
lm(formula = Percent ~ Judge - 1, data = case0502.df)
```

Coefficients:

| JudgeA | JudgeB | JudgeC       | JudgeD |
|--------|--------|--------------|--------|
| 34.12  | 33.62  | 29.10        | 27.00  |
| JudgeE | JudgeF | JudgeSpock's |        |
| 26.97  | 26.80  | 14.62        |        |

```

> ## Typical regression summary (beware R^2 and F-test):
> summary(case0502cell.lm)

```

Call:

```
lm(formula = Percent ~ Judge - 1, data = case0502.df)
```

Residuals:

| Min     | 1Q     | Median | 3Q    | Max    |
|---------|--------|--------|-------|--------|
| -17.320 | -4.367 | -0.250 | 3.319 | 14.780 |

```
Coefficients:
```

|              | Estimate | Std. Error | t value | Pr(> t )     |
|--------------|----------|------------|---------|--------------|
| JudgeA       | 34.120   | 3.092      | 11.034  | 1.47e-13 *** |
| JudgeB       | 33.617   | 2.823      | 11.909  | 1.45e-14 *** |
| JudgeC       | 29.100   | 2.305      | 12.626  | 2.35e-15 *** |
| JudgeD       | 27.000   | 4.889      | 5.523   | 2.38e-06 *** |
| JudgeE       | 26.967   | 2.823      | 9.553   | 9.18e-12 *** |
| JudgeF       | 26.800   | 2.305      | 11.628  | 3.02e-14 *** |
| JudgeSpock's | 14.622   | 2.305      | 6.344   | 1.72e-07 *** |

```
---
```

```
Signif. codes:
```

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 6.914 on 39 degrees of freedom

Multiple R-squared: 0.9486, Adjusted R-squared: 0.9394

F-statistic: 102.9 on 7 and 39 DF, p-value: < 2.2e-16

```
> ## ATYPICAL! ANOVA table:
```

```
> anova(case0502cell.lm)
```

Analysis of Variance Table

Response: Percent

|           | Df | Sum Sq | Mean Sq | F value | Pr(>F)        |
|-----------|----|--------|---------|---------|---------------|
| Judge     | 7  | 34432  | 4918.9  | 102.89  | < 2.2e-16 *** |
| Residuals | 39 | 1864   | 47.8    |         |               |

```
---
```

```
Signif. codes:
```

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> ## For later reference:
```

```
> sum(fitted(case0502cell.lm)^2)
```

```
[1] 34432.29
```

In class, we will answer the following questions about the previous Chunk's output in the context of matrices. You may want to look back at Lecture 6 to help you answer these questions.

- What does the  $\mathbf{R} \mathbf{X}$  matrix look like? (See previous output.)
- What are the LS estimators/estimates of the  $\mu_i$ ?
- What is the estimator/estimate of the variance,  $\sigma^2$  (or of the standard deviation,  $\sigma$ )? Given that our model is just a linear regression model on  $p = a = 7$  dummy variables, one for each level of Judge, we can use Definition 6.3 for MSE (or root thereof), and realize that it is given by `summary` here in the same way as we've seen for linear regression. (Again, we're essentially doing the same things whether we call them regressions or ANOVAs, though, as we said, this may be a bit simplistic.)
- What are the estimated standard errors of the estimators of the  $\mu_i$ ?
- R gives default t-tests for each of the parameters associated with the Judge factor and assumes a null value of zero by default. How are these tests computed? Are these tests interesting?
- How are the p-values for the above tests computed?
- What are the remaining quantities in the output of the `summary` function? (TO BE SURE: The  $R^2$  and adjusted  $R^2$  in the R output are not what we would expect, which you would know if you read `help(lm)`, which talks about what happens when we (R) omit the column of 1's from  $\mathbf{X}$ , as we did here. Also, the F-test is not what we might expect, either! More below.)

While R often “automatically” reports an overall F-test for equality of means via `summary` or `anova` (with only one fitted model), the omission of the intercept in this example causes R to instead report an overall F-test for all parameters (cell means) being zero. Is this interesting? More in class. In order to get the usual ***overall F-test*** (§7.9) for the equality of the Judge means, we can fit a reduced model that says that all means are *equal* (but does not say what they’re equal to), then compute the F-test in R using the

`anova` function (§7.9.1). Alternatively, we can construct a particular set of linear combinations of the means ( $\mathbf{C}\boldsymbol{\beta}$ ) to test for equality of means. We'll look at both the F vs. R approach (§6.7.4 and 7.9.1) and linear combinations approach (§6.7.3 and 7.9.2) to this overall F-test in R, in the next two chunks, then discuss on the board in class what is going on.

Be prepared to write additional notes!

First, a “Full vs. Reduced” (aka “extra sums of squares”) approach to the overall F-test of equal cell/factor level/treatment level means:

```
> ## Reduced model for overall F-test of equal means.
> case0502R.lm<- lm(Percent ~ 1, data=case0502.df)
> summary(case0502R.lm)
```

Call:

```
lm(formula = Percent ~ 1, data = case0502.df)
```

Residuals:

| Min      | 1Q      | Median | 3Q     | Max     |
|----------|---------|--------|--------|---------|
| -20.1826 | -6.6326 | 0.9174 | 5.7924 | 22.3174 |

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )   |
|-------------|----------|------------|---------|------------|
| (Intercept) | 26.583   | 1.353      | 19.64   | <2e-16 *** |

---

Signif. codes:

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 9.179 on 45 degrees of freedom

```
> ## Usual F-test for equal means via F v R approach:
> anova(case0502R.lm, case0502cell.lm)
```

Analysis of Variance Table

| Model 1: Percent ~ 1 | Model 2: Percent ~ Judge - 1 |    |           |   |        |
|----------------------|------------------------------|----|-----------|---|--------|
| Res.Df               | RSS                          | Df | Sum of Sq | F | Pr(>F) |

```

1      45 3791.5
2      39 1864.4  6     1927.1 6.7184 6.096e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Or, use a particular set of linear combinations  $\mathbf{C}\beta$  (contrasts in this case) to test the equality of means:

```

> ## Or, test the linear combination...
> library(gmodels)
> Cmat<- matrix(c(1, -1, 0, 0, 0, 0,
+                  0, 1, -1, 0, 0, 0,
+                  0, 0, 1, -1, 0, 0,
+                  0, 0, 0, 1, -1, 0,
+                  0, 0, 0, 0, 1, -1,
+                  0, 0, 0, 0, 0, 1), ncol=7, byrow=TRUE)
> b0<- rep(0,6) ## why 6?
> glh.test(case0502cell.lm, cm=Cmat, d=b0)

```

```

Test of General Linear Hypothesis
Call:
glh.test(reg = case0502cell.lm, cm = Cmat, d = b0)
F = 6.7184, df1 = 6, df2 = 39, p-value = 6.096e-05

```

### 9.3 Cell Means Model: Further Inference About Means

As in regression, typically, after examining the (overall F) question of equality of means (no linear association with the covariates), we may typically proceed to infer about other, more interesting linear combination(s) of means i.e., about other  $\mathbf{C}\beta$ , as we will generally want to do in the case of higher-way ANOVA.  $\mathbf{C}\beta$ . In a very real sense, as we've said, there is nothing new here compared to linear regression, except, perhaps, that we have to be a bit more aware of the interpretation of parameters in order to sensibly implement inferences about the parameters, of course! For now,  $\beta$  is just a vector of cell means, which is about as easy as things get; just be sure that you know which means go with which factor levels!

### 9.3.1 Example

Here we continue the Spock Conspiracy Trial example introduced above. Of particular interest in the trial was whether the Spock judge was biased against including women in his venires (panels or pools of potential jurors). Thus, it is natural to test whether the mean percent women in the Spock judge's venires is different (or perhaps less than) the mean percent women on the remaining judge's venires. That is, we are interested in comparing

$$\mu_{SpockJudge} \text{ vs } \frac{\mu_A + \mu_B + \mu_C + \mu_D + \mu_E + \mu_F}{6}$$

- What is the  $\beta$  vector (as R sees it)?
- What is an appropriate  $\mathbf{C}$  (row) matrix for estimating/testing this comparison?
- Can you present a test/estimate in a well organized manner and use R to carry out the test/estimate?
- We'll discuss this more in class, along with the implementation in R presented in the next code chunk.
- Again, be prepared to write notes in class.

```
> ## We assume existence of objects from previous code chunks.
>
> ## First, let us convince ourselves of the order in which R sees the
> ## Judge factor levels.
> levels(case0502.df$Judge)

[1] "A"          "B"          "C"          "D"          "E"
[6] "F"          "Spock's"

> case0502cell.lm
```

```
Call:  
lm(formula = Percent ~ Judge - 1, data = case0502.df)
```

## Coefficients:

| JudgeA | JudgeB | JudgeC       | JudgeD |
|--------|--------|--------------|--------|
| 34.12  | 33.62  | 29.10        | 27.00  |
| JudgeE | JudgeF | JudgeSpock's |        |
| 26.97  | 26.80  | 14.62        |        |

```

> ## Now, let's proceed to test  $H_0: CB=d$ :
> Cmat<- c(1,1,1,1,1,1,-6)/6 ## (not really an R matrix)
> b0<- 0
> ##
>
> ## Use glh.test...
> glh.test(reg=case0502cell.lm, cm=Cmat, d=b0)

```

## Test of General Linear Hypothesis

Call:

```
glh.test(reg = case0502cell.lm, cm = Cmat, d = b0)
F = 32.1459, df1 = 1, df2 = 39, p-value = 1.489e-06
```

> ##...or use *estimable* (*ugly printout*)

```
> (my.est<- estimable(obj=case0502cell.1m, cm=Cmat, beta0=d, conf.int=0.95))
```

(0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666666)

(0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666666)

(0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666666)

(0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666666)

(0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666666)

(0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666667 0.1666666666666666)

(0 16666666666666667 0 16666666666666667 0 16666666666666667 0 16666666666666667 0 16666666666666667)

(0, 16666666666666667, 0, 16666666666666667, 0, 16666666666666667, 0, 16666666666666667, 0, 16666666666666667)

```
> ## Change row.name attribute for nicer printout:  
> attr(x=my.est, which="row.names")<- ""  
> round(my.est, 3)  
  
beta0 Estimate Std. Error t value DF Pr(>|t|) Lower.CI  
0 14.978 2.642 5.67 39 0 9.635  
Upper.CI  
20.322
```

- What is the estimate of the above linear combination (contrast) (in terms of matrices)?
- What is the estimated standard error (in terms of matrices)? (What's  $\hat{\sigma}$ ?)
- How are the test statistics formed (F or t)?
- How are the p-values computed?
- How is the confidence interval computed?
- Conclusion/Interpretation/Scope of Inference?

## 9.4 Factor Effects Parameterization

Here, we continue our introduction to one-way ANOVA, now using a different parameterization of the cell means. We'll see the multi-way ANOVA version of this parameterization later, where it is more commonly used, but seeing it here may help us gain familiarity. See [KNNL05, Sec. 16.7, 16.8]; I don't recall that [RS13] discuss factor effects in the context of one-way ANOVA. See [Wak13, Expr. (5.48)]; he mentions the sum-to-zero constraint/coding, but proceeds with corner-point constraint (treatment or cell-reference), which we'll get to, shortly.

### 9.4.1 Defining a Factor Effects Parameterization Using Treatment Means

Referring to the cell means model, above, we define

$$\mu_{\cdot} = \frac{\sum_{i=1}^a \mu_i}{a},$$

which is commonly referred to as the **overall (constant) mean (effect)**, and define

$$\tau_i = (\mu_i - \mu_{\cdot}) \quad i = 1, \dots, a,$$

which is commonly referred to as the **effect** of the  $i$ th factor (treatment) level (constant for each factor level). Now we can re-write our one-way ANOVA model as

$$Y_{ij} \stackrel{\text{ind}}{\sim} N(\mu_{\cdot} + \tau_i, \sigma^2) \quad i = 1, \dots, a, \quad j = 1, \dots, n_i$$

or, equivalently,

$$Y_{ij} = \mu_{\cdot} + \tau_i + \epsilon_{ij} \quad \epsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad i = 1, \dots, a, \quad j = 1, \dots, n_i,$$

where all else is as defined before. (We'll discuss this in class, of course.)

Notice that

$$\mu_i = E(Y_{ij}) = \mu_{\cdot} + \tau_i$$

so that we have essentially the **same mean model but different parameters** with different interpretations than cell means. We have reparameterized the cell means model, which, again, is commonly referred to as an (factor) effects parameterization. Below, we will refer to another (factor) effects parameterization, but we will distinguish the two by reference to different constraints on the parameters.

## 9.5 Factor Effects Parameterization: Before Constraints

Let's reconsider the factor effects parameterization, above, but, now, we do not define its parameters in terms of cell means,  $\mu_i$ . Instead, we define a factor effects model, then define cell means in terms of the factor effects parameters. In other words,

$$Y_{ij} = \mu_+ + \tau_i + \epsilon_{ij} \quad \epsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad i = 1, \dots, a, \quad j = 1, \dots, n_i,$$

and define

$$\mu_i = E(Y_{ij}) = \mu_+ + \tau_i.$$

Obviously, the parameter  $\mu_i$  has the interpretation of a cell mean!

Now (without the definitions of  $\mu_+$  and  $\tau_i$ , as in the previous section (in terms of cell means)!), we have redundancy in our model parameters, i.e., our mean model is currently **overparameterized** (notice we have one more parameter than the cell means model). Another way to say this is that our mean model parameters are **not identifiable** or **not estimable**. One way to see this redundancy or non-identifiability is by adding and subtracting the same constant to the mean model (aka adding zero!). For example,

$$\begin{aligned} E(Y_{ij}) &= (\mu_+ + 5.73) + (\tau_i - 5.73) \\ &= \mu_*^* + \tau_i^*. \end{aligned}$$

In other words, after adding/subtracting an arbitrary constant, we end up with the same model! (Certainly, adding stars to parameter names does not change our model!) In other words, our model cannot help us to identify particular values for  $\mu_+$  and  $\tau_i$ , i.e., without a constraint on the parameters, the interpretation of the individual parameters is arbitrary. (Note that their sum is identified; it's the  $i$ th cell mean, of course.) **A bit more in class.**

Yet another way to see this redundancy is by considering the (now, non-identifiable) parameter vector

$$\boldsymbol{\beta} = (\mu_+, \tau_1, \dots, \tau_a)^T$$

and its associated  $\mathbf{X}$  matrix, constructed with a column of 1's for the constant,  $\mu_.$ , and  $a$  (number or factor levels) columns of 0/1's, one for each  $\tau_i$ . **To be done in class.** We'll call the construction of the columns of  $\mathbf{X}$ , using these 0/1 values, “indicator coding,” (or “dummy coding”) for reasons to be explained more in classes. (It may already be obvious to you because we discussed this type of coding for the cell means model.)

This redundant, indicator coding is used here to set us up for working with R. In particular, let's look at how we (very (most?)) frequently in practice) specify a one-way ANOVA model formula in R (unlike our cell means model formula, previously).

```
> ## Using the Spock Trial Data as an e.g.:
> my.lm<- lm(Percent ~ Judge, data=case0502.df)
```

Again, the model formula,

$$\text{Percent} \sim \text{Judge},$$

is by far the most frequent way to specify a one-way ANOVA (at least in R). (Our illustration of the cell means model, above, using “-1,” is not typical.) [RS13] adopt a similar “shorthand notation” for specifying mean (regression function) models with categorical variables (factors); see [RS13, Sec. 9.3.5] (SLR), [RS13, pg 390] and various other places throughout their book (their shorthand uses all caps for factors, e.g., JUDGE).

To understand what R does, it may help to envision  $\text{Percent} \sim \text{Judge}$  as denoting the redundant “indicator coding”  $\mathbf{X}$  matrix, discussed above, and its associated overparameterized  $\beta$  vector, remembering that R includes a column of 1's in the  $\mathbf{X}$  matrix, by default.

So, how does R resolve the redundancy issue? Read on.

## 9.6 Imposing Constraints

There are several ways (that R uses) to remove redundancy amongst the columns of the  $\mathbf{X}$  matrix, above. Each way makes the (remaining) parameters identifiable and gives them a particular interpretation. It's probably a good idea that you know how to interpret the parameters in your own model!

Or, at least, if you're going to use a statistical package, it's a good idea to reconcile your understanding of your model parameters with how they may be tweaked by the statistical package you're using!

## 9.7 Sum-to-Zero Constraint/Coding

One (common) resolution is to impose the constraint  $\sum_{i=1}^a \tau_i = 0$ . Thus, our model is

$$Y_{ij} = \mu. + \tau_i + \epsilon_{ij} \quad \epsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad i = 1, \dots, a, \quad j = 1, \dots, n_i,$$

now including the (sum-to-zero) constraint,

$$\sum_{i=1}^a \tau_i = 0.$$

Notice that this way of defining the mean model in the effects parameterization, with sum-to-zero constraint, is equivalent to our reparameterization of the cell means model, above, in terms of the overall mean,  $\mu.$  and deviations from the overall mean  $\tau_i = \mu_i - \mu.$  (Section 9.4.1).

How? We have

$$\mu_i = \mu. + \tau_i,$$

so that taking the mean of the  $\mu_i$  shows that

$$\mu.$$

is the **overall mean** (not just some arbitrary constant effect in a non-identifiable model), as when we first defined an effects model using means, above, and

$$\tau_i = \mu_i - \mu.$$

is the deviation of the  $i$ th treatment mean from the overall mean, not just a deviation from some arbitrary constant level,  $\mu..$  **A bit more in class.**

With the sum-to-zero constraint, we have (e.g.)

$$\tau_a = - \sum_{i=1}^{a-1} \tau_i.$$

That is, we can work with (e.g.) the first  $(a - 1)$   $\tau_i$  parameters (and the parameter  $\mu$ ) and solve for  $\tau_a$  if we need to.

- What is  $\beta$  in this case? (more in class)
- What is  $\mathbf{X}$  in this case? (more in class)
- We'll see how to implement this constraint/ $\mathbf{X}$  coding in R in a straightforward manner.

Again, the interpretation of parameters in this factor effects parameterization **with** sum-to-zero constraints is the same as in the factor effects reparameterization as originally defined above (Section 9.4.1).

The next chunk shows how to implement the sum-to-zero constraint for the continuing example using the Spock Trial Data. By default, R implements “treatment coding” or “cell reference coding” (as we’ve seen in §7.13; and, see next section) for all factors that do not have a “contrast” attribute set. There are (at least) two ways to change how R treats coding the  $\mathbf{X}$  matrix for factors.

```
> ## Does the Judge factor have a contrasts attribute?
> attr(x=case0502.df$Judge, which="contrasts")
NULL

> attributes(x=case0502.df$Judge)
$levels
[1] "A"      "B"      "C"      "D"      "E"
[6] "F"      "Spock's"

$class
[1] "factor"

> ## No. (Unless you've been goofing with the code!)
>
> ## We can change the constraint/coding for _all_ factors
> ## that do not have their own contrasts attribute:
>getOption("contrasts")
```

```
[1] "contr.treatment" "contr.treatment"

> options(contrasts = rep("contr.sum", 2))
>getOption("contrasts")

[1] "contr.sum" "contr.sum"

> ## Or, we can assign a constraint/coding to a particular factor
> ## (which will override the value of option("contrasts")):
> contrasts(case0502.df$Judge)<- contr.sum(levels(case0502.df$Judge))
> attr(x=case0502.df$Judge, which="contrasts")

[,1] [,2] [,3] [,4] [,5] [,6]
A      1     0     0     0     0     0
B      0     1     0     0     0     0
C      0     0     1     0     0     0
D      0     0     0     1     0     0
E      0     0     0     0     1     0
F      0     0     0     0     0     1
Spock's -1    -1    -1    -1    -1    -1
```

Now that we've specified the coding, we can repeat our previous cell means analysis now using the factor effects parameterization with sum-to zero constraint/coding, which, as we said above, corresponds to the factor effects model of Section 9.4.1 wherein we have the interpretation of an overall mean effect and **deviations from that overall mean** (not from some arbitrary constant or from the mean of a reference level (shortly)).

```
> case0502sum.lm<- lm(Percent ~ Judge, data=case0502.df)
> ## Let's look at the X matrix along side Judge:
> cbind.data.frame(model.matrix(case0502sum.lm),
+                   Judge=substr(case0502.df$Judge,1,1))

(Intercept) Judge1 Judge2 Judge3 Judge4 Judge5 Judge6
1           1     -1     -1     -1     -1     -1     -1
2           1     -1     -1     -1     -1     -1     -1
3           1     -1     -1     -1     -1     -1     -1
4           1     -1     -1     -1     -1     -1     -1
5           1     -1     -1     -1     -1     -1     -1
6           1     -1     -1     -1     -1     -1     -1
7           1     -1     -1     -1     -1     -1     -1
```

|       |   |    |    |    |    |    |    |
|-------|---|----|----|----|----|----|----|
| 8     | 1 | -1 | -1 | -1 | -1 | -1 | -1 |
| 9     | 1 | -1 | -1 | -1 | -1 | -1 | -1 |
| 10    | 1 | 1  | 0  | 0  | 0  | 0  | 0  |
| 11    | 1 | 1  | 0  | 0  | 0  | 0  | 0  |
| 12    | 1 | 1  | 0  | 0  | 0  | 0  | 0  |
| 13    | 1 | 1  | 0  | 0  | 0  | 0  | 0  |
| 14    | 1 | 1  | 0  | 0  | 0  | 0  | 0  |
| 15    | 1 | 0  | 1  | 0  | 0  | 0  | 0  |
| 16    | 1 | 0  | 1  | 0  | 0  | 0  | 0  |
| 17    | 1 | 0  | 1  | 0  | 0  | 0  | 0  |
| 18    | 1 | 0  | 1  | 0  | 0  | 0  | 0  |
| 19    | 1 | 0  | 1  | 0  | 0  | 0  | 0  |
| 20    | 1 | 0  | 1  | 0  | 0  | 0  | 0  |
| 21    | 1 | 0  | 0  | 1  | 0  | 0  | 0  |
| 22    | 1 | 0  | 0  | 1  | 0  | 0  | 0  |
| 23    | 1 | 0  | 0  | 1  | 0  | 0  | 0  |
| 24    | 1 | 0  | 0  | 1  | 0  | 0  | 0  |
| 25    | 1 | 0  | 0  | 1  | 0  | 0  | 0  |
| 26    | 1 | 0  | 0  | 1  | 0  | 0  | 0  |
| 27    | 1 | 0  | 0  | 1  | 0  | 0  | 0  |
| 28    | 1 | 0  | 0  | 1  | 0  | 0  | 0  |
| 29    | 1 | 0  | 0  | 1  | 0  | 0  | 0  |
| 30    | 1 | 0  | 0  | 0  | 1  | 0  | 0  |
| 31    | 1 | 0  | 0  | 0  | 1  | 0  | 0  |
| 32    | 1 | 0  | 0  | 0  | 0  | 1  | 0  |
| 33    | 1 | 0  | 0  | 0  | 0  | 1  | 0  |
| 34    | 1 | 0  | 0  | 0  | 0  | 1  | 0  |
| 35    | 1 | 0  | 0  | 0  | 0  | 1  | 0  |
| 36    | 1 | 0  | 0  | 0  | 0  | 1  | 0  |
| 37    | 1 | 0  | 0  | 0  | 0  | 1  | 0  |
| 38    | 1 | 0  | 0  | 0  | 0  | 0  | 1  |
| 39    | 1 | 0  | 0  | 0  | 0  | 0  | 1  |
| 40    | 1 | 0  | 0  | 0  | 0  | 0  | 1  |
| 41    | 1 | 0  | 0  | 0  | 0  | 0  | 1  |
| 42    | 1 | 0  | 0  | 0  | 0  | 0  | 1  |
| 43    | 1 | 0  | 0  | 0  | 0  | 0  | 1  |
| 44    | 1 | 0  | 0  | 0  | 0  | 0  | 1  |
| 45    | 1 | 0  | 0  | 0  | 0  | 0  | 1  |
| 46    | 1 | 0  | 0  | 0  | 0  | 0  | 1  |
| Judge |   |    |    |    |    |    |    |
| 1     | S |    |    |    |    |    |    |
| 2     | S |    |    |    |    |    |    |

|    |   |
|----|---|
| 3  | S |
| 4  | S |
| 5  | S |
| 6  | S |
| 7  | S |
| 8  | S |
| 9  | S |
| 10 | A |
| 11 | A |
| 12 | A |
| 13 | A |
| 14 | A |
| 15 | B |
| 16 | B |
| 17 | B |
| 18 | B |
| 19 | B |
| 20 | B |
| 21 | C |
| 22 | C |
| 23 | C |
| 24 | C |
| 25 | C |
| 26 | C |
| 27 | C |
| 28 | C |
| 29 | C |
| 30 | D |
| 31 | D |
| 32 | E |
| 33 | E |
| 34 | E |
| 35 | E |
| 36 | E |
| 37 | E |
| 38 | F |
| 39 | F |
| 40 | F |
| 41 | F |
| 42 | F |
| 43 | F |
| 44 | F |

```
45      F
46      F
```

```
> ## Estimated coefficients are given by the default printout:
> case0502sum.lm
```

Call:

```
lm(formula = Percent ~ Judge, data = case0502.df)
```

Coefficients:

|             | Judge1  | Judge2  | Judge3 |
|-------------|---------|---------|--------|
| (Intercept) | 27.4608 | 6.1559  | 1.6392 |
| Judge4      | 6.6592  | 6.1559  | 1.6392 |
| -0.4608     | -0.4941 | -0.6608 |        |

```
> ## Typical regression summary:
```

```
> summary(case0502sum.lm)
```

Call:

```
lm(formula = Percent ~ Judge, data = case0502.df)
```

Residuals:

| Min     | 1Q     | Median | 3Q    | Max    |
|---------|--------|--------|-------|--------|
| -17.320 | -4.367 | -0.250 | 3.319 | 14.780 |

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )   |
|-------------|----------|------------|---------|------------|
| (Intercept) | 27.4608  | 1.1547     | 23.781  | <2e-16 *** |
| Judge1      | 6.6592   | 2.8571     | 2.331   | 0.0250 *   |
| Judge2      | 6.1559   | 2.6504     | 2.323   | 0.0255 *   |
| Judge3      | 1.6392   | 2.2644     | 0.724   | 0.4734     |
| Judge4      | -0.4608  | 4.2903     | -0.107  | 0.9150     |
| Judge5      | -0.4941  | 2.6504     | -0.186  | 0.8531     |
| Judge6      | -0.6608  | 2.2644     | -0.292  | 0.7720     |

---

Signif. codes:

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 6.914 on 39 degrees of freedom

Multiple R-squared: 0.5083, Adjusted R-squared: 0.4326

F-statistic: 6.718 on 6 and 39 DF, p-value: 6.096e-05

```
> ## Typical ANOVA table:  
> anova(case0502sum.lm)  
  
Analysis of Variance Table  
  
Response: Percent  
          Df Sum Sq Mean Sq F value    Pr(>F)  
Judge       6 1927.1  321.18  6.7184 6.096e-05 ***  
Residuals  39 1864.5   47.81  
---  
Signif. codes:  
 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- What does the R  $\mathbf{X}$  matrix look like? (see output!)
- What are the LS estimators/estimates of the parameters?
- How do we interpret these (estimated) parameters?
- How do these estimates compare to those given in the chunk beginning on page 339 for the cell means model?
- What is the estimator/estimate of the variance,  $\sigma^2$  (or of the standard deviation,  $\sigma$ )? How does this compare with the estimate given by the cell means analysis on page 339?
- What are the estimated standard errors of the estimators of the parameters?
- R gives default t-tests for each of the parameters assuming a null value of zero by default. How are these tests computed? Are these tests interesting?
- How are the p-values for the above tests computed?
- What are the remaining quantities in the output of the **summary** function? Unlike the automatic overall F-test given with the cell means

analysis in the chunk on page 339, the overall F-test here *is* for equality of means (all  $\tau_i = 0$  but no test for overall mean (associated with the 1's covariate as is the overall intercept in linear regression, which is not part of the overall F-test null hypothesis). (Compare to our F v. R approach or linear combinations approach for testing the equality of means back in the cell means section. We'll verify (in class) that these approaches give the same overall F-test results as we get here; see below. Again, same regression model, different parameterization, same overall F-test.)

- Scope of inference doesn't change, of course!

Though we have a “good” overall F-test for equality of means given in the output of the `anova` function (and from `summary`), above, still we mimic our cell means presentation above, by showing the F v. R approach to the overall F-test for equality of means as well as showing the linear combinations approach, which have slight changes here, compared to the cell means case, above, to accomodate (illustrate) the factor effect parameterization with sum-to-zero constraint/coding.

Be prepared to write additional notes!

First, a “Full vs. Reduced” (or “extra sums of squares”) approach to the overall F-test of equal cell means. This is essentially the same implementation as our cell mean analysis, above.

```
> ## Reduced model (same as before) for overall F-test of equal means.
> case0502R.lm<- lm(Percent ~ 1, data=case0502.df)
> summary(case0502R.lm)
```

Call:  
`lm(formula = Percent ~ 1, data = case0502.df)`

Residuals:

```

      Min       1Q   Median      3Q      Max
-20.1826 -6.6326  0.9174  5.7924 22.3174

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.583     1.353   19.64 <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.179 on 45 degrees of freedom

> ## Usual F-test for equal means via F v R approach:
> anova(case0502R.lm, case0502sum.lm)

Analysis of Variance Table

Model 1: Percent ~ 1
Model 2: Percent ~ Judge
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     45 3791.5
2     39 1864.4  6    1927.1 6.7184 6.096e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Or, use a set of linear combinations (set of contrasts) to test the equality of treatment means. Notice that our matrix  $\mathbf{C}$  is different here than for the cell means analysis, above! We have different  $\beta$  vectors, in these two cases, right?! Again, know your parameterization! More in class.

```

> ## Or, test the set of linear combinations...
> library(gmodels)
> Cmat<- matrix(c(0, 1, 0, 0, 0, 0, 0,
+                  0, 0, 1, 0, 0, 0, 0,
+                  0, 0, 0, 1, 0, 0, 0,
+                  0, 0, 0, 0, 1, 0, 0,
+                  0, 0, 0, 0, 0, 1, 0,
+                  0, 0, 0, 0, 0, 0, 1), ncol=7, byrow=TRUE)
> b0<- rep(0,6)
> glh.test(case0502sum.lm, cm=Cmat, d=b0)

```

```
Test of General Linear Hypothesis
Call:
glh.test(reg = case0502sum.lm, cm = Cmat, d = b0)
F = 6.7184, df1 = 6, df2 = 39, p-value = 6.096e-05
```

## 9.8 Reference Treatment Constraint/Coding

Another (common) way to resolve redundancy in the  $\mathbf{X}$  matrix of the effects parameterization is to impose the constraint  $\tau_1 = 0$  (as in R; [Wak13, p. 225]). Thus, our model is

$$Y_{ij} = \mu. + \tau_i + \epsilon_{ij} \quad \epsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad i = 1, \dots, a, \quad j = 1, \dots, n_i,$$

now including the constraint,

$$\boxed{\tau_1 = 0}.$$

Thus, we work the remaining  $\tau_i$ ,  $i = 2, \dots, a$  (and  $\mu.$ ) and the associated non-redundant (full column rank)  $\mathbf{X}$ .

Notice that our parameters have the same names/symbols as in our factor effects model with the sum-to-zero constraint, but the **treatment constraint**, here, results in different interpretations of parameters in the two cases. In other words, the factor effects parameterization with treatment coding/constraint gives parameters that have different interpretations than those in the factor effects parameterization with sum-to-zero coding/constraint!

- In particular, the mean of the observations within factor level 1 ( $E(Y_{1j})$ ) is  $\mu. + 0$  so that  $\mu.$  is now the mean of the observations associated with the first (“reference”) factor level in the same way that the intercept,  $\beta_0$  was when discussing this coding in §7.13. It’s not the overall mean effect as with the sum-to-zero constraint/coding. Note that such a “reference treatment” may be a placebo treatment to which all other “active” treatments are compared. Below, I make

the Spock judge level the reference “treatment” to illustrate how this can simplify things (after showing how not doing so may be less simple); see the code chunk in Section 9.9, below.)

- The mean of the observations within factor level  $i \neq 1$  is  $\mu_+ + \tau_i$  so that  $\tau_i$ ,  $i = 2, \dots, a$ , are the effects of being associated with the  $i$ th treatment level *relative* to the first (reference) treatment level. They are not deviations from the overall mean effect as with the sum-to-zero constraint!
- This constraint is associated with its particular coding of  $\mathbf{X}$  (wonder of wonders, “cell reference coding” or “treatment coding”) which refers to the particular values (numerical codes) that arise in the resulting  $\mathbf{X}$  matrix by using this constraint.
- Note to SAS users: SAS (at one time, at least) sets the last level (not first) parameter to zero:  $\tau_a = 0$ .

- What is  $\beta$  in this case? (more in class)
- What is  $\mathbf{X}$  in this case? (more in class)
- We’ll see how to implement this constraint/ $\mathbf{X}$  coding in R in a straightforward manner.

Continuing a pattern: The next chunk shows how to implement the treatment constraint/coding for the continuing example using the Spock Trial Data. By default, R implements “treatment coding” or “cell reference coding” for all factors that do not have a “contrast” attribute set. There are (at least) two ways to change how R treats coding for factors. (Again, we’re following the cell means and sum-to-zero analyses, above, now in terms of treatment coding/constraint.)

```

> ## Does the Judge factor have a contrasts attribute?
> attr(x=case0502.df$Judge, which="contrasts")

      [,1] [,2] [,3] [,4] [,5] [,6]
A       1     0     0     0     0     0
B       0     1     0     0     0     0
C       0     0     1     0     0     0
D       0     0     0     1     0     0
E       0     0     0     0     1     0
F       0     0     0     0     0     1
Spock's -1    -1    -1    -1    -1    -1

> attributes(x=case0502.df$Judge)

$levels
[1] "A"          "B"          "C"          "D"          "E"
[6] "F"          "Spock's"

$class
[1] "factor"

$contrasts
      [,1] [,2] [,3] [,4] [,5] [,6]
A       1     0     0     0     0     0
B       0     1     0     0     0     0
C       0     0     1     0     0     0
D       0     0     0     1     0     0
E       0     0     0     0     1     0
F       0     0     0     0     0     1
Spock's -1    -1    -1    -1    -1    -1

> ## Yes, if previous chunk objects still exist. Else, no.
>
> ## We can change the constraint/coding for _all_ factors
> ## that do not have their own contrasts attribute:
>getOption("contrasts")

[1] "contr.sum" "contr.sum"

> options(contrasts = rep("contr.treatment", 2))
>getOption("contrasts")

[1] "contr.treatment" "contr.treatment"

```

```
> ## Or, we can assign a constraint/coding to a particular factor
> ## (which will override the value of options("contrasts")):
> contrasts(case0502.df$Judge)<- contr.treatment(levels(case0502.df$Judge))
> attr(x=case0502.df$Judge, which="contrasts")

      B C D E F Spock's
A    0 0 0 0 0      0
B    1 0 0 0 0      0
C    0 1 0 0 0      0
D    0 0 1 0 0      0
E    0 0 0 1 0      0
F    0 0 0 0 1      0
Spock's 0 0 0 0 0      1
```

Now that we've specified the coding, we can repeat our previous cell means and sum-to-zero analyses, now using the treatment coding, which, as we said, results in different interpretations of parameters compared to the factor effects parameterization with sum-to-zero coding in Section 9.4.1, despite our using the same parameter names/symbols.

```
> case0502trmt.lm<- lm(Percent ~ Judge, data=case0502.df)
>
> ## Let's look at the X matrix along side Judge:
> tmp<- cbind.data.frame(model.matrix(case0502trmt.lm),
+                         Judge=case0502.df$Judge)
> names(tmp)<- c("(Intercept)","B","C","D","E","F","Spock's","Judge")
> tmp

  (Intercept) B C D E F Spock's Judge
1           1 0 0 0 0 0      1 Spock's
2           1 0 0 0 0 0      1 Spock's
3           1 0 0 0 0 0      1 Spock's
4           1 0 0 0 0 0      1 Spock's
5           1 0 0 0 0 0      1 Spock's
6           1 0 0 0 0 0      1 Spock's
7           1 0 0 0 0 0      1 Spock's
8           1 0 0 0 0 0      1 Spock's
9           1 0 0 0 0 0      1 Spock's
10          1 0 0 0 0 0      0      A
11          1 0 0 0 0 0      0      A
12          1 0 0 0 0 0      0      A
```

|    |             |   |   |
|----|-------------|---|---|
| 13 | 1 0 0 0 0 0 | 0 | A |
| 14 | 1 0 0 0 0 0 | 0 | A |
| 15 | 1 1 0 0 0 0 | 0 | B |
| 16 | 1 1 0 0 0 0 | 0 | B |
| 17 | 1 1 0 0 0 0 | 0 | B |
| 18 | 1 1 0 0 0 0 | 0 | B |
| 19 | 1 1 0 0 0 0 | 0 | B |
| 20 | 1 1 0 0 0 0 | 0 | B |
| 21 | 1 0 1 0 0 0 | 0 | C |
| 22 | 1 0 1 0 0 0 | 0 | C |
| 23 | 1 0 1 0 0 0 | 0 | C |
| 24 | 1 0 1 0 0 0 | 0 | C |
| 25 | 1 0 1 0 0 0 | 0 | C |
| 26 | 1 0 1 0 0 0 | 0 | C |
| 27 | 1 0 1 0 0 0 | 0 | C |
| 28 | 1 0 1 0 0 0 | 0 | C |
| 29 | 1 0 1 0 0 0 | 0 | C |
| 30 | 1 0 0 1 0 0 | 0 | D |
| 31 | 1 0 0 1 0 0 | 0 | D |
| 32 | 1 0 0 0 1 0 | 0 | E |
| 33 | 1 0 0 0 1 0 | 0 | E |
| 34 | 1 0 0 0 1 0 | 0 | E |
| 35 | 1 0 0 0 1 0 | 0 | E |
| 36 | 1 0 0 0 1 0 | 0 | E |
| 37 | 1 0 0 0 1 0 | 0 | E |
| 38 | 1 0 0 0 0 1 | 0 | F |
| 39 | 1 0 0 0 0 1 | 0 | F |
| 40 | 1 0 0 0 0 1 | 0 | F |
| 41 | 1 0 0 0 0 1 | 0 | F |
| 42 | 1 0 0 0 0 1 | 0 | F |
| 43 | 1 0 0 0 0 1 | 0 | F |
| 44 | 1 0 0 0 0 1 | 0 | F |
| 45 | 1 0 0 0 0 1 | 0 | F |
| 46 | 1 0 0 0 0 1 | 0 | F |

```
> ## Estimated coefficients given by default:  
> case0502trmt.lm
```

Call:  
lm(formula = Percent ~ Judge, data = case0502.df)

Coefficients:

```
(Intercept) JudgeB JudgeC JudgeD
34.1200     -0.5033    -5.0200   -7.1200
JudgeE       JudgeF  JudgeSpock's
-7.1533     -7.3200    -19.4978

> ## Typical regression summary:
> summary(case0502trmt.lm)

Call:
lm(formula = Percent ~ Judge, data = case0502.df)

Residuals:
    Min      1Q  Median      3Q      Max 
-17.320  -4.367  -0.250   3.319   14.780 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.1200    3.0921  11.034 1.47e-13 ***
JudgeB      -0.5033    4.1868  -0.120  0.9049    
JudgeC      -5.0200    3.8566  -1.302  0.2007    
JudgeD      -7.1200    5.7848  -1.231  0.2258    
JudgeE      -7.1533    4.1868  -1.709  0.0955 .  
JudgeF      -7.3200    3.8566  -1.898  0.0651 .  
JudgeSpock's -19.4978   3.8566  -5.056 1.05e-05 ***

---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.914 on 39 degrees of freedom
Multiple R-squared:  0.5083, Adjusted R-squared:  0.4326 
F-statistic: 6.718 on 6 and 39 DF,  p-value: 6.096e-05

> ## Typical ANOVA table:
> anova(case0502trmt.lm)

Analysis of Variance Table

Response: Percent
          Df Sum Sq Mean Sq F value    Pr(>F)    
Judge      6 1927.1 321.18  6.7184 6.096e-05 ***
Residuals 39 1864.5 47.81
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- What does the **R X** matrix look like? (see output!)
- What are the LS estimators/estimates of the parameters?
- How do we interpret these (estimated) parameters?
- How do these estimates compare to those given in the chunk beginning on page 339 for the cell means model? For the sum-to-zero (factor effects model) on given in the chunk beginning on page 353?
- What is the estimator/estimate of the variance,  $\sigma^2$  (or of the standard deviation,  $\sigma$ )? How does this compare to the estimate given by the cell means analysis in page 339? The that given by the sum-to-zero (factor effects model) on given in the chunk beginning on page 353?
- What are the estimated standard errors of the estimators of the parameters?
- R gives default t-tests for each of the parameters associated with the **Judge** factor and assumes a null value of zero by default. How are these tests computed? Are these tests interesting?
- How are the p-values for the above tests computed?
- What are the remaining quantities in the output of the **summary** function? Unlike the automatic overall F-test given with the cell means analysis in the chunk on page 339, the overall F-test here *is* for equality of means (all  $\tau_i = 0$ ). (Compare to our F v. R approach or linear combinations approach for testing the equality of means back in the cell means sections and in the sum-to-zero analysis sections. We'll verify that these approached give the same overall F-test results here, too; see below.)
- Again, scope of inference does not change!

Though we have a “good” overall F-test for equality of means given in

the output above (as we did for the sum-to-zero analysis but not for our cell means analysis) still we mimic our cell means and sum-to-zero presentations above, by showing the F v. R approach to the overall F-test for equality of means and the linear combinations approach, which have slight changes here to acknowledge the sum-to-zero constraint/coding.

Be prepared to write additional notes!

First, a “Full vs. Reduced” (or “extra sums of squares”) approach to the overall F-test of equal cell means. This is essentially the same implementation as our cell means and sum-to-zero analyses, above.

```
> ## Reduced model (same as before) for overall F-test of equal means.  
> case0502R.lm<- lm(Percent ~ 1, data=case0502.df)  
> summary(case0502R.lm)
```

Call:  
lm(formula = Percent ~ 1, data = case0502.df)

Residuals:  
Min 1Q Median 3Q Max  
-20.1826 -6.6326 0.9174 5.7924 22.3174

Coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 26.583 1.353 19.64 <2e-16 \*\*\*  
---

Signif. codes:  
0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.179 on 45 degrees of freedom

```
> ## Usual F-test for equal means via F v R approach:  
> anova(case0502R.lm, case0502trmt.lm)
```

Analysis of Variance Table

```
Model 1: Percent ~ 1  
Model 2: Percent ~ Judge
```

```

Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1       45 3791.5
2       39 1864.4  6     1927.1 6.7184 6.096e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Or, use a set of linear combinations (set of contrasts) to test the equality of means. Notice that our matrix **C** is different here than for the cell means analysis but the same as for sum-to-zero analysis, above! Why? More in class.

```

> ## Or, test the set of linear combinations...
> library(gmodels)
> Cmat<- matrix(c(0, 1, 0, 0, 0, 0,
+                  0, 0, 1, 0, 0, 0,
+                  0, 0, 0, 1, 0, 0,
+                  0, 0, 0, 0, 1, 0,
+                  0, 0, 0, 0, 0, 1),
+                  ncol=7, byrow=TRUE)
> b0<- rep(0,6)
> glh.test(case0502trmt.lm, cm=Cmat, d=b0)

```

```

Test of General Linear Hypothesis
Call:
glh.test(reg = case0502trmt.lm, cm = Cmat, d = b0)
F = 6.7184, df1 = 6, df2 = 39, p-value = 6.096e-05

```

## 9.9 Further Inference About Treatment Means

As we did for the cell means model of  $E(Y_{ij})$  (Section 9.3), we can make further, more detailed inferences about treatment means in the factor effects parameterization using either the sum-to-zero constraint or the treatment constraint. (Recall §6.7). We merely need to change our implementation to accomodate the particular coding. Again, we'll repeat this sort of analysis when we get to higher-way ANOVA later.

### 9.9.1 Example

Again, we continue the Spock Conspiracy Trial example. Recall, of particular interest in the trial was whether the Spock judge was biased against including women in his venires (panels of potential jurors). Thus, it is natural to test whether the mean percent women in the Spock judge's venires is different (or perhaps less than) the mean percent women on the remaining judge's venires. That is we are interested in comparing

$$\mu_{SpockJudge} \text{ vs } \frac{\mu_A + \mu_B + \mu_C + \mu_D + \mu_E + \mu_F}{6}.$$

More specifically, we want to infer the linear combination (contrast),

$$\frac{\mu_A + \mu_B + \mu_C + \mu_D + \mu_E + \mu_F}{6} - \mu_{SpockJudge}.$$

For the factor effects parameterization (using either the sum-to-zero constraint or the treatment constraint), we recall that

$$\mu_i = \mu_{\cdot} + \tau_i,$$

which, expressing the linear combination in these terms, gives (after a bit of algebra perhaps)

$$\frac{\tau_A + \tau_B + \tau_C + \tau_D + \tau_E + \tau_F}{6} - \tau_{SpockJudge}.$$

For the sum-to-zero constraint, we have, in addition,  $\sum_i \tau_i = 0$  so that, recalling  $SpockJudge$  ( $i = a = 7$ ) to be the last factor level,

$$\tau_{SpockJudge} = - \sum_{i=1}^{a-1} \tau_i,$$

and the linear combination can be written in terms of the (non-redundant) sum-to-zero parameters,

$$\frac{7}{6}(\tau_A + \tau_B + \tau_C + \tau_D + \tau_E + \tau_F),$$

which we can use in the sum-to-zero analysis, below.

For the treatment constraint, we have, instead,  $\tau_1 = 0$  ( $\tau_A = 0$ ), whereby the linear combination can be expressed in terms of its (non-redundant) parameters as

$$\frac{\tau_B + \tau_C + \tau_D + \tau_E + \tau_F}{6} - \tau_{SpockJudge},$$

which we can use in the treatment coding analysis, below.

As this example illustrates, the implementation of inference about linear combinations of means is intuitive when using the cell means model. Still, the re-expression in terms of the effects model (without constraints) is not so bad. This is particularly true for constraints where constants such as  $\mu$ . drop out, leaving a relatively intuitive expression in terms of the  $\tau_i$ , as illustrated here. But, accommodating constraints takes (only) a bit of work to arrive at, perhaps, an expression that may not be so intuitive. In other words, you might want to revert to the cell means model to infer linear combinations of treatment means. Still, once we have our linear combinations expressed in the appropriate parameterization (and constraints), as we just derived, we can proceed with the inference. As we see in the chunk below and from the corresponding chunk in Section 9.3, inferences are the same! This is as we should expect because the factor effects parameterization (either coding) is just a reparameterization (alternative representation) of the same cell means model.

Notice that, in the treatment constraint/coding, we did not make the Spock judge level the reference level. In the R code below, I do this to show it makes the implementation of our comparison a bit easier to understand. Again, the reference level is often chosen for convenience of interpretation/understanding.

```
> ## We assume existence of objects from previous code chunks.
>
> ## First, let double check how R views factor levels:
> levels(case0502.df$Judge)
[1] "A"        "B"        "C"        "D"        "E"
[6] "F"        "Spock's"

> ## Sum-to-zero parameterization:
> ##
```

```

> ## Ho: CB=d:
> Cmat<- c(0,7,7,7,7,7,7)/6
> b0<- 0
> ##
> ## Use glh.test...
> glh.test(reg=case0502sum.lm, cm=Cmat, d=b0)

Test of General Linear Hypothesis
Call:
glh.test(reg = case0502sum.lm, cm = Cmat, d = b0)
F = 32.1459, df1 = 1, df2 = 39, p-value = 1.489e-06

> ##...or use estimable (avoiding ugly printout)
> my.est<- estimable(obj=case0502sum.lm, cm=Cmat, beta0=d, conf.int=0.95)
> attr(x=my.est, which="row.names")<- ""
> round(my.est, 3)

beta0 Estimate Std. Error t value DF Pr(>|t|) Lower.CI
0 14.978 2.642 5.67 39 0 9.635
Upper.CI
20.322

> ## Treatment parameterization:
> ##
> ## Ho: CB=d:
> Cmat<- c(c(0,1,1,1,1,1)/6, -1)
> b0<- 0
> ##
> ## Use glh.test...
> glh.test(reg=case0502trmt.lm, cm=Cmat, d=b0)

Test of General Linear Hypothesis
Call:
glh.test(reg = case0502trmt.lm, cm = Cmat, d = b0)
F = 32.1459, df1 = 1, df2 = 39, p-value = 1.489e-06

> ##...or use estimable (avoiding ugly printout)
> my.est<- estimable(obj=case0502trmt.lm, cm=Cmat, beta0=d, conf.int=0.95)
> attr(x=my.est, which="row.names")<- ""
> round(my.est, 3)

```

```

beta0 Estimate Std. Error t value DF Pr(>|t|) Lower.CI
      0    14.978     2.642     5.67 39       0    9.635
Upper.CI
      20.322

> ## Note how making the Spock's level the reference level makes
> ## interpretation a bit more natural. (Assuming here that the
> ## contrast attribute is still set to contr.treatment, which
> ## should be the case unless you've messed with the data/analysis!)
> (tmp<- levels(case0502.df$Judge))

[1] "A"        "B"        "C"        "D"        "E"
[6] "F"        "Spock's"

> case0502.df$Judge<- factor(case0502.df$Judge,
+                                levels=tmp[c(7,1:6)])
> levels(case0502.df$Judge)

[1] "Spock's"  "A"        "B"        "C"        "D"
[6] "E"        "F"

> tmp.lm<- lm(Percent ~ Judge, data=case0502.df)
> Cmat<- c(0,1,1,1,1,1,1)/6 ## a bit more transparent...
> b0<- 0 ##...yes? no?
> glh.test(reg=tmp.lm, cm=Cmat, d=b0)

Test of General Linear Hypothesis
Call:
glh.test(reg = tmp.lm, cm = Cmat, d = b0)
F = 32.1459, df1 = 1, df2 = 39, p-value = 1.489e-06

> ## BTW, above tests if mean (singular) of others is same as Spock
> ## mean, which is different than testing all others' means (plural)
> ## are equal to Spock mean, which is just the overall F-test for
> ## equality of all means, which doesn't change if we change reference
> ## levels (though it's implementation via a linear combinations approach
> ## may require changes).
> anova(tmp.lm)

Analysis of Variance Table

Response: Percent

```

```
Df Sum Sq Mean Sq F value    Pr(>F)
Judge      6 1927.1  321.18  6.7184 6.096e-05 ***
Residuals 39 1864.5   47.81
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As in the cell means Section 9.3, we should be comfortable with discussing the estimator  $\mathbf{C}\hat{\beta}$ , its standard error, t/F statistics, interval estimator for  $\mathbf{C}\beta$  (when  $\mathbf{C}$  is a row matrix), p-value, and presentation in a way that demonstrates command of the material. **Perhaps more in class.**

## 9.10 Summary of One-Way ANOVA

So what is ‘ANalysis Of VAriance (ANOVA)?’ Despite having just presented a fair amount of material on ANOVA, this question is a bit premature in the one-way case. But, we can begin to answer it here. We’ll return to it briefly in the multi-way ANOVA case, later.

In regression, ANOVA and linear models in general, we decompose the variability of the observations into two overall pieces; “overall” in the sense that we may be able to decompose things further, but we will only consider further decomposition, later, in multi-way ANOVA with equal number of observations per treatment. Using our notation for observations,  $Y_{ij}$  (in one-way ANOVA context), we can write

$$\sum_{ij} (Y_{ij} - \bar{Y})^2 = \sum_{ij} (\hat{Y}_{ij} - \bar{Y})^2 + \sum_{ij} (Y_{ij} - \hat{Y}_{ij})^2,$$

which is almost universally denoted as

$$\text{SSTO} = \text{SSTR} + \text{SSE},$$

where

- SSTO is “(corrected) total sum-of-squares,” a measure of the variability of the observations about their overall (sample) average,  $\bar{Y}$ .

“Corrected” refers to having subtracted (or “corrected for”) the (sample) average before squaring. (We’re typically not interested in the variability of observations about zero.) This is somehow a measure of the total variability in our observations. (Note that it is the numerator of the sample variance, a popular summary of the variability of observed values.)

- SSTR is “treatment sum-of-squares,” a measure of the variability of the estimated mean values,  $\hat{Y}_{ij}$  (fitted values), about the overall (sample) average, i.e., a measure of the response variability captured by the estimated mean model, i.e., a measure of the variability in the response associated with the covariates (factors). For one-way ANOVA,  $\hat{Y}_{ij} = \hat{\mu}_i$ , using the cell means model, for example. [RS13] tend to refer to SSTR as “Between group” sum-of-squares, which is suggestive of SSTR being a measure of the variability in observations that is associated with treatment groups.
- SSE is “error sum-of-squares,” a measure of the variability of the observations about their estimated mean (fitted) values, i.e., the residual variability after accounting for SSTR in SSTO. [RS13] tend to refer to SSE as “within group” sum-of-squares, which is suggestive of SSE being a measure of the variability in observations that is, well, within groups.
- [Wak13] attempts to give this decomposition for the general linear model in the middle of [Wak13, p. 213], but there are typos there. He uses notation TSS = FSS + RSS, corresponding to our SSTO = SSTR + SSE. We touched on this simple decomposition of the total sums of squares using his notation when discussing the overall F-test in §7.9 and  $R^2$  in §7.10. [Wak13, Tab. 5.5] gets the decomposition right. The LS fitted values  $\hat{Y}_{ij} = \mathbf{x}_i^t \hat{\beta} = \bar{Y}_i$  in his table, just the empirical average of observations in level  $i$ . In his table, SSTR is the first sum of squares, SSE is the second, and SSTO is the last. We saw a table like this for the Spock example from the

`anova` function, above, which looks a lot like like [Wak13, Tab. 5.6] (different example).

We also have an analogous (and obvious) decomposition of the total degrees of freedom:

$$(n_T - 1) = (a - 1) + (n_T - a),$$

where

- $(n_T - 1)$  is the “total degrees of freedom”
- $(a - 1)$  is the “treatment degrees of freedom” and
- $(n_T - a)$  is the “error degrees of freedom” (or “residual degrees of freedom”).

We will not go into how these degrees of freedom arise, but will say that they are used for statistical inference (wonder of wonders) by dividing each into their respective sums-of-squares (to get “mean squares” MSTO, MSTR and MSE, although MSTO is not typically used and these mean squares are not additive like the SS’s and df’s). We will say more in class in a sort of passing manner.

This decomposition is typically reported in an “ANOVA table.” In R, it’s given by `anova` applied to your linear model fit (to the object that is returned by `lm`; we’ve seen this already, above). In addition to sums-of-squares, we usually see (some of) the corresponding mean squares. We almost never see MSTO (sample variance), and we often do not see SSTO in ANOVA tables (you can add to get it). We also usually see one or more F statistics and p-values. In the one-way case, with one factor, we only see one mean one mean square, MSTR, in addition to MSE.

Look back at the chunks where `anova` is applied to fitted linear model (`lm`) objects to give ANOVA tables. **In class, we will reconcile the R ANOVA**

tables with our discussion here. As mentioned before, the ANOVA table for the cell means case is atypical; its reported SSTR and associated degrees of freedom do not account for the overall sample mean. (This has to do with having removed the column of 1's from  $\mathbf{X}$  in R, and is apparent also in the atypical F-test given by the `summary` function applied to the fitted cell means model object.)

In short, ANOVA tables simply summarize test(s) that may be of interest. Here, in one-way ANOVA, we saw only one “overall F-test” which we performed using the F vs. R approach and the linear combinations approach, so we really didn’t have much of a need for an ANOVA table, did we?

Forewarning: In the multi-way ANOVA case, with unequal sample sizes per treatment (so called unbalanced data), ANOVA tables may potentially be confusing. But, our F vs. R (or linear combinations) approach will continue to serve us well. More later.

### 9.10.1 Regression Approach to ANOVA

Notice that imposing constraints led us to an  $\mathbf{X}$  matrix that is full rank, just like in regression (usually), whereby we can get the usual LS solution for  $\boldsymbol{\beta}$  as we’ve discussed before ( $\mathbf{X}'\mathbf{X}$  is invertible). Each column of the resulting  $\mathbf{X}$  matrix can be thought of as the observations of a regression variable (with specially coded values). Such coding of regression variables is discussed in textbooks under such headings as “Regression Approach to ANOVA” or something similar. See., e.g., [KNNL05, Sec. 16.7, 16.8] for an illustration using the effects parameterization with sum-to-zero constraint/coding. [RS13] wait until two-way ANOVA to discuss the “regression approach to ANOVA.”

### 9.10.2 Model, Parametrization, Reparameterization

We should note that we have treated only one model in this section in the sense that the cell means model, the factor effects (re)parameterization with sum-to-zero constraint and the factor effects (re)parameterization with treatment constraint all give the same fitted values (their  $\mathbf{X}\hat{\boldsymbol{\beta}}$  are the same); more technically, their  $\mathbf{X}$  matrices span the same column space. Less technically, we’ve seen the necessary consequences of the equivalence of these mean mod-

els: same MSE ( $\hat{\sigma}^2$ ), same ANOVA tables (notwithstanding strangeness arising from omitting the 1's column), same overall F-test results, same results for inferences of treatments means.

We started with the cell means “model.” We then relabeled the cell means model parameter symbols with new parameter symbols in a re-parameterization of the cell means model, imposing two different constraints leading to two different interpretations of “effects”. But, as mentioned, these (re)parameterizations of the cell means model are the same model. We could have started with a factor effects model with a particular constraint, then produced a cell means (re)parameterization. In any case, we might use “model” or “(re)parameterization” interchangeably to refer to the cell means model / parameterization or factor effects model / parameterization (with whichever constraint).



# Lecture 10

## Multi-Way ANOVA

### Contents

---

|   |            |
|---|------------|
| <b>10.1 Initial Concepts and Notation . . . . .</b>                             | <b>381</b> |
| <b>10.2 ANOVA Model Components: Means and Effects . . . . .</b>                 | <b>383</b> |
| <b>10.3 Example . . . . .</b>   | <b>390</b> |
| <b>10.4 Cell Means Model of <math>E(Y_{ijk})</math> . . . . .</b>               | <b>395</b> |
| <b>10.5 Factor Effects Parameterization: Before Constraints . . . . .</b>       | <b>396</b> |
| <b>10.6 Factor Effects: Sum-to-Zero Constraints/Coding . . . . .</b>            | <b>400</b> |
| 10.6.1 E.g.: Factor Effects S2Zero Initial Analysis . . . . .                   | 404        |
| 10.6.2 E.g.: Effects S2Zero ANOVA For Common $\mathbf{C}\beta$ . . . . .        | 419        |
| 10.6.3 E.g.: Factor Effects S2Zero F v R & $\mathbf{C}\beta$ Approach . . . . . | 423        |
| 10.6.4 E.g.: Effects 2Zero Summary . . . . .                                    | 425        |
| <b>10.7 Factor Effects: Treatment Constraints/Coding . . . . .</b>              | <b>425</b> |
| 10.7.1 E.g.: Factor Effects Trmt Initial Analysis . . . . .                     | 430        |
| 10.7.2 E.g.: Effects Trmt ANOVA For Common $\mathbf{C}\beta$ . . . . .          | 435        |
| 10.7.3 E.g.: Factor Effects Trmt F v R & $\mathbf{C}\beta$ Approach . . . . .   | 436        |
| 10.7.4 E.g.: Effects Trmt Summary . . . . .                                     | 438        |
| <b>10.8 SS Type, Balance &amp; the Marginality Principle . . . . .</b>          | <b>438</b> |
| 10.8.1 Sequential SS ANOVA . . . . .  | 440        |
| 10.8.2 Partial SS ANOVA . . . . .   | 443        |
| 10.8.3 Marginality Principle . . . . .  | 444        |
| 10.8.4 Balance . . . . .  | 448        |
| <b>10.9 Additive Model: Tests for Overall Main Effects . . . . .</b>            | <b>449</b> |
| 10.9.1 F v R Approach . . . . .   | 449        |
| 10.9.2 $\mathbf{C}\beta$ Approach . . . . .                                     | 452        |

**10.10 Additive Model: More Detailed Inference of Main Effects . . . 453**

---

***Main Objectives:***

- Overall tests and tests for main/interaction effects using ANOVA tables or using F v R (extra SS) approach or using  $\mathbf{C}\boldsymbol{\beta}$  approach
- Inference for (more detailed) linear combinations of  $\boldsymbol{\beta}$
- Effects model with sum-to-zero constraints or treatment constraints
- Types of SS in ANOVA tables (sequential or type I; partial or type III) (`drop1`; `car` package's `Anova` function)
- More about the marginality principle
- More about balance

---

 $\mathcal{O}$

***Additional Reading:***

[RS13, Chap. 13]

[KNNL05, Chap. 19, 23 & 24]

Topics that we do not cover directly but for which the current material is very closely related (FYI only): no treatment replication ([RS13, Chap. 14], [KNNL05, Chap. 20]); RCBD ([KNNL05, Chap. 21]); ANCOVA ([KNNL05, Chap. 22])

We give more detailed references in the sections below  $\dots \mathcal{R}$

## 10.1 Initial Concepts and Notation

Now we consider more than one factor variable in our ANOVA models. We will illustrate with two factors, whereby we may envision observations or their means to be arranged in a two-way table with rows defined by the levels of one factor, and columns are defined by the levels of the second factor. Extension to three or more factors should be obvious—tables become arrays, etc. We repeat and adapt some terminology presented previously in the one-way case.

- **Number of factor variables:** 2 (or more): Generically factor A, factor B, etc.
- **Number of factor levels**  $a$  for factor A,  $b$  for factor B, etc.
- **Treatments.** As before, treatments are the set of conditions defined by the unique combinations of factors levels across all factors. *Now that we have two (or more) factors, treatments are no longer synonymous with factor levels.*
- **Sample sizes** (or number of units per treatment level) are denoted  $n_{ij}$ ,  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ . We can sum over  $j$  to get the number

of units,  $n_{i\cdot}$ , for factor A, level  $i$ , or sum over  $i$  to get the number of units,  $n_{\cdot j}$ , for factor B, level  $j$ . Notation for, say, a three-way layout?

- **Observation**  $\boxed{Y_{ijk}}$  is the kth observation (response) at the treatment level defined by the ith level of factor A and the jth level of factor B,  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ ,  $k = 1, \dots, n_{ij}$ . Notation for, say, a three-way layout?
- **Total number of observations**, i.e., total sample size, is denoted as  $\boxed{n_T}$ , i.e.,  $n_T = n_{11} + n_{12} + \dots + n_{ab} = \sum_{i=1}^a \sum_{j=1}^b n_{ij}$ . Three-way layout?
- **Balance.** If we have an equal number of observations per treatment level, then we say our treatment design is balanced. Otherwise it is unbalanced. i.e.,  $n_{11} = n_{12} = \dots = n_{ab} = n$ , where  $n$  is the common number of observations in each treatment. Thus, in the balanced case,  $n_T = nab$ . Three-way layout? NOTE: Balance has historically received more attention than we (and [RS13]) give it.
- **Observational or experimental treatment (cell) level means** are denoted by  $\boxed{\mu_{ij}}$ ,  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ . Three-way layout?
- **Factor level (marginal) means**

$$\begin{aligned}\mu_{i\cdot} &= \sum_{j=1}^b \mu_{ij}/b \\ \mu_{\cdot j} &= \sum_{i=1}^a \mu_{ij}/a\end{aligned}$$

- Overall mean

$$\begin{aligned}
 \mu_{..} &= \sum_{i=1}^a \sum_{j=1}^b \mu_{ij} / (ab) \\
 &= \sum_{j=1}^b \mu_{.j} / b \\
 &= \sum_{i=1}^a \mu_{i.} / a
 \end{aligned}$$

- Scope of Inference See our previous, one-way presentation in §9.1.

## 10.2 ANOVA Model Components: Means and Effects

Given the above notation, we illustrate the components of two-way ANOVA models using material adapted from [KNNL05, Sec. 19.2] in the next few pages (scanned from hand written notes). We introduce notation somewhat loosely to facilitate concepts, here, but we will revisit the notation in a more formal manner, shortly. In the process, we introduce components of both the **cell means model** and **factor effects model, with sum-to-zero constraints**, in anticipation of more formal treatments of the means and effects models, in subsequent sections. We add coverage of treatment constraints, later, for an overall presentation that is similar to our previous presentation of one-way ANOVA.

Much discussion in class!

## Two Way ANOVA (equal sample sizes $n_{ij} = n$ )

**TABLE 19.1**  
Age Effect but  
No Gender  
Effect, with No  
Interactions—  
Learning  
Example.

Theoretical Example  
Does (mean) time  
to learn a task  
depend on gender?  
on age?

Recall Cell Means Model

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

$$i=1 \dots a$$

$$j=1 \dots b$$

$$k=1 \dots n_{ij}$$

we're just looking  
at cell means  
and effects for now

|                   |        | (a) Mean Learning Times (in minutes) |                             |                         |                    |
|-------------------|--------|--------------------------------------|-----------------------------|-------------------------|--------------------|
|                   |        | Factor B   Age                       |                             |                         |                    |
| Factor A   Gender |        | $j = 1$<br>Young                     | $j = 2$<br>Middle           | $j = 3$<br>Old          | Row Average        |
| $i = 1$           | Male   | 9 ( $\mu_{11}$ )                     | Cell mean 11 ( $\mu_{12}$ ) | $\mu_{13}$              | 16 ( $\mu_{13}$ )  |
| $i = 2$           | Female | 9 ( $\mu_{21}$ )                     | mean 11 ( $\mu_{22}$ )      | means 16 ( $\mu_{23}$ ) | 12 ( $\mu_{23}$ )  |
| Column average    |        | 9 ( $\mu_{..1}$ )                    | 11 ( $\mu_{..2}$ )          | 16 ( $\mu_{..3}$ )      | 12 ( $\mu_{..3}$ ) |
|                   |        | marginal means                       |                             |                         | overall mean       |

| (b) Main Gender Effects (in minutes)              |  | (c) Main Age Effects (in minutes)                |  |
|---|--|--|--|
| $\alpha_1 = \mu_{1.} - \mu_{..} = 12 - 12 = 0$    |  | $\beta_1 = \mu_{.1} - \mu_{..} = 9 - 12 = -3$    |  |
| $\alpha_2 = \mu_{2.} - \mu_{..} = 12 - 12 = 0$    |  | $\beta_2 = \mu_{.2} - \mu_{..} = 11 - 12 = -1$   |  |
| $\alpha_i = \mu_{i.} - \mu_{..}$ Factor A Effects |  | $\beta_j = \mu_{.j} - \mu_{..}$ Factor B Effects |  |

Definitions:

Marginal Means

$$A \quad \mu_{i.} = \sum_{j=1}^b \mu_{ij} / b$$

$$B \quad \mu_{.j} = \sum_{i=1}^a \mu_{ij} / a$$

$$\text{Overall Mean} \quad \mu_{..} = \sum_{i=1}^a \sum_{j=1}^b \mu_{ij} / (ab)$$

check for yourself!

This page illustrates additive effects (not always the case!!!)

(\*) More on Effects

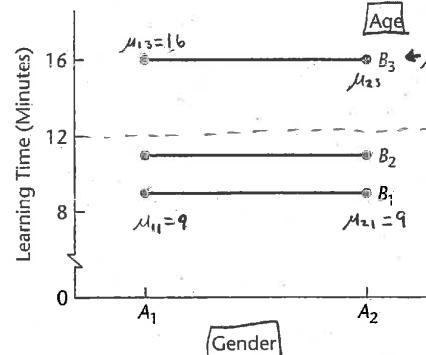
Definitions result in sum to zero:

$$\sum_{i=1}^a \alpha_i = 0$$

$\sum_{j=1}^b \beta_j = 0$   
These will become constraints on our 2-way factor effects model (coming soon)

**FIGURE 19.3**  
Age Effect but  
No Gender  
Effect, with No  
Interactions—  
Learning  
Example.

"Interaction" plot



$$\begin{aligned} \mu_{13} &= 16 \\ B_3 &\leftarrow \mu_{03} = \frac{\mu_{13} + \mu_{23}}{2} = 16 \\ \mu_{23} &= 12 \\ \mu_{..} &= 12 \\ (\text{in this } \epsilon_{ij},) \quad \mu_1 &= \mu_2 = 12 \\ \Rightarrow \alpha_i &= 12 - 12 = 0 \\ i=1,2 \end{aligned}$$

Interpretation of effects:

- $\alpha_i$  — on average (over levels of Factor B), the "blip" (effect) up/down from the overall mean  $\mu_{..}$  due to being in the  $i$ th level of Factor A
- $\beta_j$  — on average (over levels of Factor A), the "blip" (effect) up/down from the overall mean  $\mu_{..}$  due to being in the  $j$ th level of Factor B

**TABLE 19.2**  
Age and  
Gender Effects,  
with No  
Interactions—  
Learning  
Example.

|                 |        | (a) Mean Learning Times (in minutes) |                    |                    |                    |
|-----------------|--------|--------------------------------------|--------------------|--------------------|--------------------|
|                 |        | Factor B—Age                         |                    |                    |                    |
| Factor A—Gender |        | $j = 1$<br>Young                     | $j = 2$<br>Middle  | $j = 3$<br>Old     | Row<br>Average     |
| $i = 1$         | Male   | 11 ( $\mu_{11}$ )                    | 13 ( $\mu_{12}$ )  | 18 ( $\mu_{13}$ )  | 14 ( $\mu_{1..}$ ) |
| $i = 2$         | Female | 7 ( $\mu_{21}$ )                     | 9 ( $\mu_{22}$ )   | 14 ( $\mu_{23}$ )  | 10 ( $\mu_{2..}$ ) |
| Column average  |        | 9 ( $\mu_{..1}$ )                    | 11 ( $\mu_{..2}$ ) | 16 ( $\mu_{..3}$ ) | 12 ( $\mu_{...}$ ) |

| (b) Main Gender Effects (in minutes)               |  | (c) Main Age Effects (in minutes)                 |
|--|--|---|
| $\alpha_1 = \mu_{1..} - \mu_{....} = 14 - 12 = 2$  |  | $\beta_1 = \mu_{1..} - \mu_{....} = 9 - 12 = -3$  |
| $\alpha_2 = \mu_{2..} - \mu_{....} = 10 - 12 = -2$ |  | $\beta_2 = \mu_{2..} - \mu_{....} = 11 - 12 = -1$ |
|  |  | $\beta_3 = \mu_{3..} - \mu_{....} = 16 - 12 = 4$  |

**FIGURE 19.4**  
Age and  
Gender Effects,  
with No  
Interactions—  
Learning  
Example.

Tell-tale sign  
of no interaction:  
parallel lines

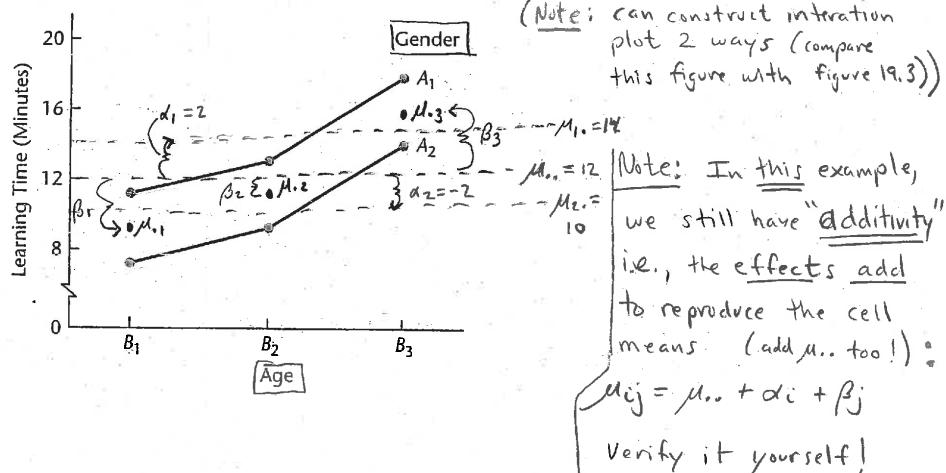


TABLE 19.3  
Age and  
Gender Effects,  
with  
Interactions—  
Learning  
Example.

|                 |        | (a) Mean Learning Times (in minutes) |                    |                    | Main<br>Gender<br>Effect |
|-----------------|--------|--------------------------------------|--------------------|--------------------|--------------------------|
|                 |        | Factor B—Age                         |                    |                    |                          |
| Factor A—Gender |        | $j = 1$<br>Young                     | $j = 2$<br>Middle  | $j = 3$<br>Old     | Row<br>Average           |
| $i = 1$         | Male   | 9 ( $\mu_{11}$ )                     | 12 ( $\mu_{12}$ )  | 18 ( $\mu_{13}$ )  | 13 ( $\mu_{1..}$ )       |
| $i = 2$         | Female | 9 ( $\mu_{21}$ )                     | 10 ( $\mu_{22}$ )  | 14 ( $\mu_{23}$ )  | 11 ( $\mu_{2..}$ )       |
| Column average  |        | 9 ( $\mu_{..1}$ )                    | 11 ( $\mu_{..2}$ ) | 16 ( $\mu_{..3}$ ) | 12 ( $\mu_{...}$ )       |
| Main age effect |        | -3 ( $\beta_1$ )                     | -1 ( $\beta_2$ )   | 4 ( $\beta_3$ )    |                          |

|                |  | (b) Interactions (in minutes) |         |         | Row<br>Average |
|----------------|--|-------------------------------|---------|---------|----------------|
|                |  | $j = 1$                       | $j = 2$ | $j = 3$ |                |
| $i = 1$        |  | -1                            | 0       | 1       | 0              |
| $i = 2$        |  | 1                             | 0       | -1      | 0              |
| Column average |  | 0                             | 0       | 0       | 0              |

All interaction effects

$(\alpha\beta)_{ij}$  must be zero before we say main effects are additive.

We have "non-additivity" of main effects, i.e., there is an interaction effect in this e.g.

### Definition

#### Interaction effects

$$(\text{one symbol!}) (\alpha\beta)_{ij} \equiv \mu_{ij} - (\mu_{..} + \alpha_i + \beta_j)$$

"extent to which the additive model cannot reproduce the cell means"

We will have the additional sum-to-zero constraints:

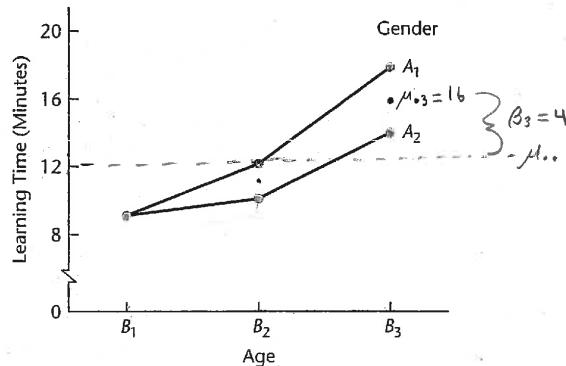
$$\sum_{i=1}^a (\alpha\beta)_{ij} = 0 = \sum_{j=1}^b (\alpha\beta)_{ij}$$

Why are these "important"?

FIGURE 19.5  
Age and  
Gender Effects,  
with Important  
Interactions—  
Learning  
Example.

Tell-tail sign  
of interaction:

non-parallel  
lines

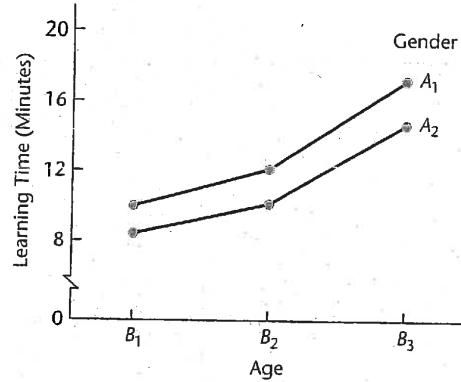


Interpretation of interaction effects: The "blip" up/down due to being in cell  $i, j$  after accounting for the average effect of Factor A level  $i$  and the average effect of Factor B level  $j$ .

**TABLE 19.4**  
Age and  
Gender Effects,  
with  
Unimportant  
Interactions—  
Learning  
Example.

|                 |        | Factor B—Age     |                   |                | Row Average |
|-----------------|--------|------------------|-------------------|----------------|-------------|
| Factor A—Gender |        | $j = 1$<br>Young | $j = 2$<br>Middle | $j = 3$<br>Old |             |
| $i = 1$         | Male   | 9.75             | 12.00             | 17.25          | 13.00       |
|                 | Female | 8.25             | 10.00             | 14.75          | 11.00       |
| Column average  |        | 9.00             | 11.00             | 16.00          | 12.00       |

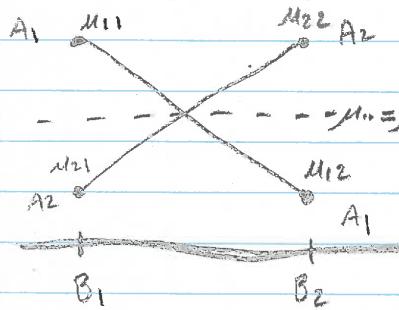
**FIGURE 19.6**  
Age and  
Gender Effects,  
with  
Unimportant  
Interactions  
(curves almost  
parallel)—  
Learning  
Example.



Try computing  
interaction effects.  
They're small here.  
Moreover, they're "unimportant"  
(why?)

An illustrative e.g. of what can happen when main effects are used to characterize differences in factor level means when "bad" interactions exists.

2 Factors A, B each at 2 levels



Obviously, something interesting is going on with the mean response between levels of A within a level of B (simples are interesting), but using main effects to look at differences in mean

response averages the simples to get a misleading answer. Point: Beware of interpreting main effects when interactions are present.

- Simple effect of A at  $B_1$ :  $\mu_{21} - \mu_{11}$

- "Simple effect" of A at  $B_2$ :  $\mu_{22} - \mu_{12}$

- Using Main Effects:

$$\alpha_2 - \alpha_1 \quad (\text{why do this?})$$

$$= (\mu_{21} - \mu_{11}) - (\mu_{22} - \mu_{12})$$

$$= (\mu_{21} - \mu_{12}) \quad (\text{OK, I see})$$

$$= \frac{1}{2} \{ \mu_{21} + \mu_{22} - (\mu_{11} + \mu_{12}) \}$$

$$= \frac{1}{2} \{ \underbrace{(\mu_{21} - \mu_{11})}_{\text{Simple}} + \underbrace{(\mu_{22} - \mu_{12})}_{\text{Simple}} \}$$

$$= 0 \quad (\text{in this example})$$

826 Part Five Multi-Factor Studies

**TABLE 19.6**  
Examples of  
Different Types  
of Interactions.

|              |  | (a) Productivity of Executives |       |
|--------------|--|--------------------------------|-------|
|              |  | Factor B—Authority             |       |
| Factor A—Pay |  | Small                          | Great |
| Low          |  | 50                             | 72    |
| High         |  | 74                             | 75    |

|              |  | (b) Productivity of Executives |       |
|--------------|--|--------------------------------|-------|
|              |  | Factor B—Authority             |       |
| Factor A—Pay |  | Small                          | Great |
| Low          |  | 50                             | 52    |
| High         |  | 53                             | 75    |

|              |  | (c) Productivity of Executives |       |
|--------------|--|--------------------------------|-------|
|              |  | Factor B—Authority             |       |
| Factor A—Pay |  | Small                          | Great |
| Low          |  | 50                             | 72    |
| High         |  | 72                             | 50    |

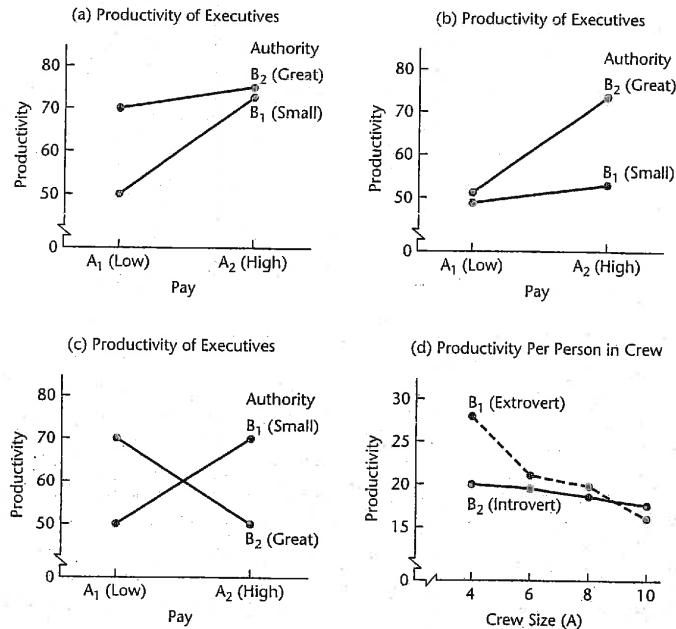
  

|                    |  | (d) Productivity per Person in Crew |           |
|--------------------|--|-------------------------------------|-----------|
|                    |  | Factor B—Personality of Crew Chief  |           |
| Factor A—Crew Size |  | Extrovert                           | Introvert |
| 4 persons          |  | 28                                  | 20        |
| 6 persons          |  | 22                                  | 20        |
| 8 persons          |  | 20                                  | 19        |

More examples for your edification.

Chapter 19 Two-Factor Studies with Equal Sample Sizes 829

**FIGURE 19.7**  
Treatment  
Means Plots—  
Examples of  
Interactions  
from  
Table 19.6.

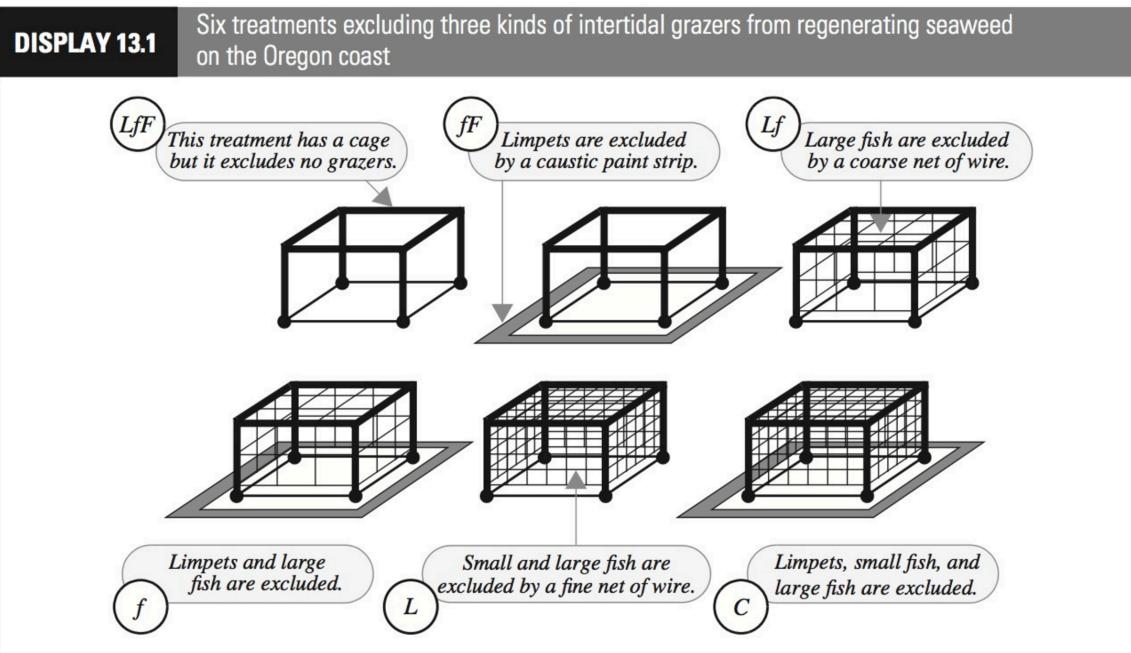


### 10.3 Example

Our next example illustrates, further, the ANOVA model components introduced above. The example is taken from [RS13, Sec. 13.1.1]. We will use it throughout much of our discussion of two-way ANOVA.

**Example 10.1** (Intertidal Seaweed Grazers). (*not used in a pejorative sense, of course*) Researchers designed an experiment to investigate the impacts of grazing on the regeneration rates of seaweed. See [RS13, Sec. 13.1.1] for detailed discussion. The basic “treatment” design (more in class) is shown in the next figure ([RS13, Display 13.1]).

The intertidal zone is a highly variable environment, which may affect regeneration, but this is (presumably) well known among such researchers, and is not of primary interest. Still, the researchers wish to account for environmental effects in an effort to help elucidate the effects of grazing, their primary interest. The researchers **replicate** their basic “treatment” design at eight locations, which we call **blocks**, with, in this case, each block (location) itself containing two replications of the basic “treatment” design. The second figure below illustrates the configuration of the observations by “treatments” and blocks ([RS13, Display 13.2]). Note that it is the combination of the “treatments” factor and the blocks factor that defines what we called “treatments,” i.e., [RS13, Sec. 13.1.1] use the term “treatments” differently than we’ve defined it. More in class.



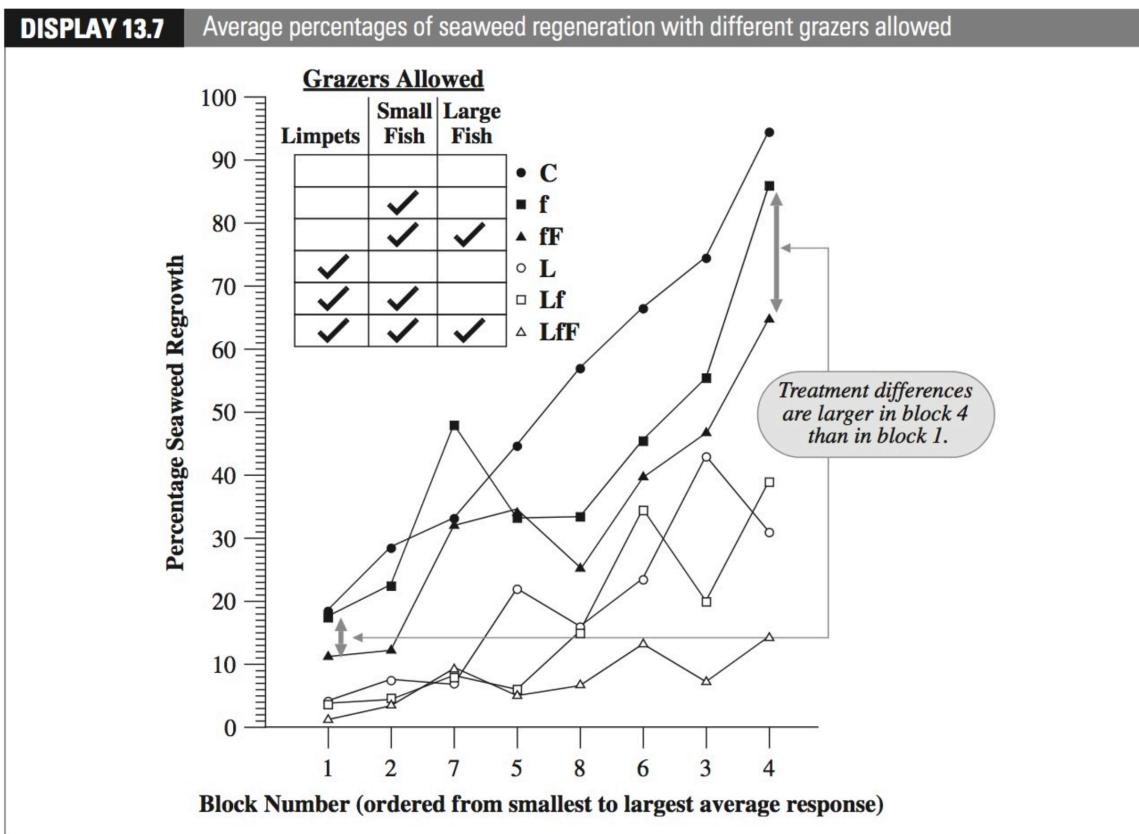
**DISPLAY 13.2**

Percentage of regenerated seaweed cover on plots with different grazers excluded, in eight blocks of differing tidal situation and exposure

| Block # | Treatment: Grazers with access |    |    |    |    |    |    |    |    |    |     |    |
|---------|--------------------------------|----|----|----|----|----|----|----|----|----|-----|----|
|         | Control                        |    | L  |    | f  |    | Lf |    | fF |    | LfF |    |
| 1       | 14                             | 23 | 4  | 4  | 11 | 24 | 3  | 5  | 10 | 13 | 1   | 2  |
| 2       | 22                             | 35 | 7  | 8  | 14 | 31 | 3  | 6  | 10 | 15 | 3   | 5  |
| 3       | 67                             | 82 | 28 | 58 | 52 | 59 | 9  | 31 | 44 | 50 | 6   | 9  |
| 4       | 94                             | 95 | 27 | 35 | 83 | 89 | 21 | 57 | 57 | 73 | 7   | 22 |
| 5       | 34                             | 53 | 11 | 33 | 33 | 34 | 5  | 9  | 26 | 42 | 5   | 6  |
| 6       | 58                             | 75 | 16 | 31 | 39 | 52 | 26 | 43 | 38 | 42 | 10  | 17 |
| 7       | 19                             | 47 | 6  | 8  | 43 | 53 | 4  | 12 | 29 | 36 | 5   | 14 |
| 8       | 53                             | 61 | 15 | 17 | 30 | 37 | 12 | 18 | 11 | 40 | 5   | 7  |

Let's reconcile this example with our concepts/definitions for two-way ANOVA before proceeding.

- What are the factors? Factor levels?  $a$ ?  $b$ ?
- What are the treatments? (Careful. The term “treatment” in the problem statement is different than we defined it.)
- $n_{ij}$ ? Balanced?
- $Y_{ijk}$ ?
- What would an interaction plot look like? See [RS13, Display 13.7], reproduced nearby.



- **Remark:** The example conjures three factors (not including blocks), each at two levels, defining  $2^3 = 8$  “treatment” levels—a three-way ANOVA or, more particularly a  **$2^3$  factorial design** (3 factors, each at 2 levels) of treatments. Again, not including blocks. See [RS13, Chap. 24] and [KNNL05, Chaps. 15 & 29]. But, consider the two levels defined by the exclusion of small fish and the inclusion of large fish, including or excluding limpets... (are you thinking?)... Thus, we consider here only  $b = 6$  levels for the “treatment” factor, and do not refer to a “factorial” structure for “treatments.” We just have a single “treatment” factor (in addition to the block/location factor). Again, don’t confuse “treatments”!
- **Remark:** Our example fits the definition of what is called a **randomized complete block design (RCBD)**. **Blocks** or **blocking**

is a relatively special experimental design topic; see [RS13, Chap. 24] and [KNNL05, Chaps. 21 & 28].) Blocking is a generalization of the notion of **pairing** or **match pairs** procedures that you may have seen in an introductory course; “pair” refers to two levels of “treatments” within each level of block, i.e., two treatment levels are somehow matched by each level of block.

- **Remark:** For us, we simply consider a two-way layout defined by two factors. We do not give much consideration to the special nature of RCBD, factorial designs, or other particular notions of experimental designs.

The next Chunk gets the data from the **Sleuth3** package and looks at them briefly. **Cover** is the response variable: percent cover of regenerated seaweed in a plot. Note that R already views the **Block** and **Treat** variables as factors. We’ll continue with these data in subsequent sections.

```
> case1301.df<- Sleuth3::case1301
> ##
> ## How is the data ``structured?:
> str(case1301.df)

'data.frame': 96 obs. of  3 variables:
 $ Cover: int  14 23 22 35 67 82 94 95 34 53 ...
 $ Block: Factor w/ 8 levels "B1","B2","B3",...: 1 1 2 2 3 3 4 4 5 5 ...
 $ Treat: Factor w/ 6 levels "C","L","Lf","LfF",...: 1 1 1 1 1 1 1 1 1 1 ...

> ## First few obs:
> head(case1301.df)

  Cover Block Treat
1     14    B1     C
2     23    B1     C
3     22    B2     C
4     35    B2     C
5     67    B3     C
6     82    B3     C
```

```
> ## Last few:
> tail(case1301.df)

  Cover Block Treat
91     10    B6   LfF
92     17    B6   LfF
93      5    B7   LfF
94     14    B7   LfF
95      5    B8   LfF
96      7    B8   LfF
```

## 10.4 Cell Means Model of $E(Y_{ijk})$

The cell means model for two-way (or higher-way) ANOVA is essentially the same as in one-way AVOVA, up to an obvious and slight change of notation. See [KNNL05, Sec. 19.3]. As before, we build a model for the mean, i.e., for the regression function, i.e., for  $E(Y_{ijk})$ , which, as before, is simple:  $E(Y_{ijk}) = \mu_{ij}$ . That is, the observations in the  $ij$ th treatment (not factor now!) level have their own mean, symbolized by a single parameter  $\mu_{ij}$  for each level.

$$Y_{ijk} \stackrel{\text{ind}}{\sim} N(\mu_{ij}, \sigma^2) \quad i = 1, \dots, a, \quad j = 1, \dots, b \quad k = 1, \dots, n_{ij}$$

or, equivalently (recall basic results in Lecture 3),

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \quad \epsilon_{ijk} \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad i = 1, \dots, a, \quad j = 1, \dots, b \quad k = 1, \dots, n_{ij}$$

where, as before,  $\epsilon_{ijk}$  is the **error** of observation  $k$  in the treatment level defined by level  $i$  of factor A and level  $j$  of factor B, and its (error) variance (component),  $\sigma^2$ , is assumed to be common to all treatment levels.

Again, we can write this cell means model in the form of a general linear model using matrix notation.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_T}).$$

As in the one-way case, above, we will write the elements of each of the vectors/matrices, above, on the board, in class. Please be prepared to take additional notes!

The cell means model typically is not the first choice for analyzing data in a multi-way layout. Still, when first choices (below) fail, in some sense, we may return to the cell means model for further analysis. This is different than for the one-way layout, where the cell means model is often a very reasonable first choice, though a factor effects model is common, too. For higher-way layouts, a factor effects model is typically employed. We do not analyze the data here in terms of the cell means model. We could, but we have limited time, we've covered it before, and use of the cell means model here, as we said, is not typical except, perhaps, in “messy data” situations. (Note, we may ignore factors and define means  $\mu_i$ ,  $i = 1, \dots, ab$ , in which case our one-way cell means model applies directly.)

## 10.5 Factor Effects Parameterization: Before Constraints

Refer to [RS13, 13.5.6] for a very brief discussion on the (additive; no interactions) factor effects model (parameterization) and to [KNNL05, Sec. 19.3] for a two-way presentation and [KNNL05, Sec. 24.1] for a three-way presentation; these presentations discuss the factor effects model with the sum-to-zero constraints, not treatment constraints, which we will discuss, later.

We've already discussed ANOVA model components, above, in Section 10.2, and we've seen the effects models for the one-way layout. The parameters in the effects models for higher way layouts have similar interpretations as before, for each of the sum-to-zero constraints and treatment constraints. We discuss these constraints in the following two sections.

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad \epsilon_{ijk} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

$$i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n_{ij}.$$

where

$$\mu_{..}$$

is an overall constant effect common to all observations in all treatments; it will be an overall mean, as discussed above, when we impose sum-to-zero constraints (as the .. subscript anticipates), but it will not be an overall mean in the case of treatment constraints, similar to the one-way case;

$$\alpha_i$$

is an effect common to all observations in the  $i$ th level of factor A;

$$\beta_j$$

is an effect common to all observations in the  $j$ th level of factor B; and

$$(\alpha\beta)_{ij}$$

is an effect of factor A level  $i$  within factor B level  $j$  (or vice-versa), over and above the  $\alpha_i$  and  $\beta_j$  effects.

As in the one-way layout, without constraints, we have redundancy, i.e., our mean model is currently **overparameterized**. And, again, another way to say this is that our mean model parameters are **not identifiable** or **not estimable**. Similar to the one-way factor effects parameterization (Section 9.5), we can illustrate this non-identifiability by adding zero to our effects model. For example,

$$\begin{aligned} E(Y_{ijk}) &= \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} \\ &= (\mu_{..} + 5) + (\alpha_i - 3) + (\beta_j - 1) + ((\alpha\beta)_{ij} - 1) \\ &= \mu^*_{..} + \alpha^*_i + \beta^*_j + (\alpha\beta)^*_{ij} \end{aligned}$$

and, certainly, relabeling or renaming the parameters from unstarred symbols to starred symbols does not change the model. So, we see that the parameters in our unconstrained mean model are not identifiable. In other words, without a constraint, the interpretation of the individual parameters is arbitrary. In short, just as in the one-way case, you should know what constraints are imposed (e.g., in R) so that you know how to interpret parameters and to implement inferences for parameters, e.g., specify  $\mathbf{C}\boldsymbol{\beta}$ .

Yet another way to see this redundancy is by considering the (non-identifiable) parameter vector

$$\boldsymbol{\beta} = (\mu_{..}, \alpha_1, \dots, \alpha_a, \beta_1, \dots, \beta_b, (\alpha\beta)_{11} \dots (\alpha\beta)_{ab})^T$$

and its associated  $\mathbf{X}$  matrix, constructed with a column of 1's for the constant,  $\mu_{..}$ ,  $a$  (number or factor levels) columns of 0/1's, one for each  $\alpha_i$ ,  $b$  (number or factor levels) columns of 0/1's, one for each  $\beta_i$ , and  $ab$  columns of 0/1's, one for each  $(\alpha\beta)_{ij}$ . **To be done in class.** We'll call the construction of the (redundant) columns of  $\mathbf{X}$ , using these 0/1 values, “indicator coding” or “dummy coding”, similar to the one-way case.

Again, this redundant, indicator coding is used here to set us up for working with R.

```
> ## Using the seaweed regeneration data as an e.g.:
> my.lm<- lm(Cover ~ Block + Treat + Block:Treat, data=case1301.df)
```

The model formula,

$$\text{Cover} \sim \text{Block} + \text{Treat} + \text{Block:Treat},$$

is again similar to the model shorthand notation of [RS13]; see [RS13, 13.3]. (Again, their shorthand uses all caps for factors, e.g., BLOCK). To help us understand R code/output, it may help to envision `Cover ~ Block + Treat + Block:Treat` as denoting how to construct  $\mathbf{X}$  with the redundant “dummy coding” associated with the overparameterized  $\boldsymbol{\beta}$  vector, above, remembering that R includes a column of 1's in the  $\mathbf{X}$  matrix, by default. The process of resolving redundancies, described below, may not be exactly how R resolves redundancies, but it does help us to understand the end result, which is the same despite R's actual redundancy resolution process.

## A Bit More on R Model Formulae

- The model formula

$$\text{R} \sim \text{A} + \text{B} + \text{A:B}$$

may be thought of most easily in terms of the redundant incidence matrix coding, discussed above. (Below, we discuss how R resolves

the redundancy by using constraints/coding, similar to the one-way case.) First, the formula tells R to include as the first column in (the redundant)  $\mathbf{X}$ , implicitly and by default, a column of 1's, which corresponds to  $\mu$ ... Then, A tells R to augment the column of 1's with an  $a$ -column incidence matrix for levels of factor A; the columns correspond to the  $\alpha_i$ ,  $i = 1, \dots, a$ . Then, B tells R to further augment with a  $b$ -column incidence matrix for levels of factor B; the columns correspond to the  $\beta_i$ ,  $i = 1, \dots, b$ . Finally, A:B tells R to include an  $ab$ -column matrix obtained by multiplying (element-wise!...Shur or Hadamard product) each column of the “A” (sub)matrix with each column the “B” (sub)matrix; the columns correspond to the  $(\alpha\beta)_{ij}$ ,  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ .

- The model formula

$$\mathbf{A} * \mathbf{B}$$

is an alternative formula specification that is equivalent to  $\mathbf{R} \sim \mathbf{A} + \mathbf{B} + \mathbf{A}:\mathbf{B}$ .

- The model formula

$$\mathbf{R} \sim (\mathbf{A} + \mathbf{B})^2$$

is yet another, equivalent an alternative specification; it tells R to include all factor terms (just A and B here) and all interaction terms up to 2-way interaction terms (just A:B here).

- What about formulas for higher-way (e.g., three-way) layouts (in the current context)?
- See [VR02, Sec. 6.2] and, in R, `help(formula)`.

Again, we can see the redundancy in such an incidence-coded  $\mathbf{X}$  matrix. And, again, how R resolves redundancies depends what constraints/coding we specify.

## 10.6 Factor Effects: Sum-to-Zero Constraints/Coding

One way to resolve the redundancy issue is by constraining sets of parameters to sum to zero, analogous to our discussion of the factor effects model with sum-to-zero constraints for one-way ANOVA in §9.5 and 9.7:

$$\begin{aligned}\sum_{i=1}^a \alpha_i &= 0 \\ \sum_{j=1}^b \beta_j &= 0 \\ \sum_{i=1}^a (\alpha\beta)_{ij} &= 0 \quad j=1, \dots, b, \\ \sum_{j=1}^b (\alpha\beta)_{ij} &= 0 \quad i=1, \dots, a.\end{aligned}$$

This suggests, e.g.,

$$\begin{aligned}\alpha_a &= -\sum_{i=1}^{a-1} \alpha_i, \\ \beta_b &= -\sum_{j=1}^{b-1} \beta_j, \\ (\alpha\beta)_{aj} &= -\sum_{i=1}^{a-1} (\alpha\beta)_{ij}, \quad j=1, \dots, b, \\ (\alpha\beta)_{ib} &= -\sum_{j=1}^{b-1} (\alpha\beta)_{ij}, \quad i=1, \dots, a.\end{aligned}$$

In other words, similar to the one-way case, we do not need to include  $\alpha_a$ ,  $\beta_b$  ( $\alpha\beta)_{aj}$  (all  $j$ ) or ( $\alpha\beta)_{ib}$  (all  $i$ ) in our model because we can solve for them. And, this tells us how to (re-)code the (non-redundant)  $\mathbf{X}$  matrix.

- What does the non-redundant  $\beta$  look like? (more in class)
- What does the non-redundant  $\mathbf{X}$  look like? (more in class) (BTW, all interaction term (re-coded) columns can be obtained by multiplying (element-wise!) each (re-coded) column associated with one factor by each (re-coded) column associated with another factor.
- Again, we are simply performing regression with specially coded “X” variables, one for each column in  $\mathbf{X}$ , and with associated coefficients having a particular interpretations.
- We’ll see how to implement these constraints/coding in R, shortly.
- Higher way effect model? (briefly in class)
- Non-additive model? (briefly in class)

Notice the interpretation of parameters in this factor effects model with sum-to-zero constraints.

- We can relate cell means and effects models parameters as

$$\mu_{ij} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij}.$$

Indeed, the (non-additive or saturated...more shortly) effects model is equivalent to the cell means model: their respective  $\mathbf{X}$  matrices span the same covariate/regressor space, each gives the same predicted values, same residuals, same estimate of  $\sigma^2$ , same overall F-test; this is similar to the one-way case. Perhaps more on this later or in an assignment.

- **Overall Effect.** Averaging  $\mu_{ij}$  over i and j, **with** sum-to-zero constraints, gives

$$\mu_{..} = \sum_i \sum_j \mu_{ij} / (ab),$$

so that  $\mu_{..}$  now is not merely some arbitrary overall effect but is the **overall (unweighted) mean** (average) of the  $\mu_{ij}$ , similar to the one-way case with sum-to-zero constraint.

- **Main Effects of Factor A.** *With* sum-to-zero constraints, averaging  $\mu_{ij}$  over  $j$  gives,

$$\mu_{i\cdot} \equiv \mu_{..} + \alpha_i$$

or

$$\alpha_i = \mu_{i\cdot} - \mu_{..},$$

so that  $\alpha_i$  is interpreted as the deviation from the overall mean due to being in the  $i$ th level of factor A. It is the **main effect** of the  $i$ th level of A in the sense that it is added to all observations in the  $i$ th level of A without regard for the observation's factor B level. We have already discussed (Section 10.2) that inferring main effects (lower order terms) in the presence of interactions (see bullet below) (higher order terms) may not be sensible (see Section 10.2)

- **Main Effects of Factor B.** *With* sum-to-zero constraints, averaging over  $i$  gives,

$$\mu_{\cdot j} \equiv \mu_{..} + \beta_j$$

or

$$\beta_j = \mu_{\cdot j} - \mu_{..},$$

so that  $\beta_j$  is interpreted as the deviation from the overall mean due to being in the  $j$ th level of factor B. It is the **main effect** of the  $j$ th level of B in the sense that it is added to all observations in the  $j$ th level of B without regard for the observation's factor A level. We have already discussed (Section 10.2) that inferring main effects (lower order terms) in the presence of interactions (see bullet below) (higher order terms) may not be sensible (see Section 10.2)

- **Interaction Effects.**

$$\begin{aligned}
 (\alpha\beta)_{ij} &\equiv \mu_{ij} - (\mu_{..} + \alpha_i + \beta_j) \\
 &= \mu_{ij} - \mu_{..} + (\mu_{i.} - \mu_{..}) + (\mu_{.j} - \mu_{..}) \\
 &= \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..},
 \end{aligned}$$

After accounting for an overall effect ( $\mu_{..}$ ) and the main (or global or average factor) effects ( $\alpha_i$ , and  $\beta_j$ ),  $(\alpha\beta)_{ij}$  is the additional effect required to reproduce the mean,  $\mu_{ij}$ . If we hold, e.g.,  $j$  constant, we see that we may interpret  $(\alpha\beta)_{ij}$  as an additional effect of the  $i$ th level of A, due to being at level  $j$  of factor B. (Or, If we hold, e.g.,  $i$  constant, we see that we may interpret  $(\alpha\beta)_{ij}$  as an additional effect of the  $j$ th level of B, due to being at level  $i$  of factor A.) In other words, the effects of factor A (B) depends on the level of factor B (A), and we say that the factors **interact** and call  $(\alpha\beta)_{ij}$  an **interaction effect**. Again, we've discussed interaction (Section 10.2), and we have warned about the potential nonsensicality of inferring main effects (lower order terms) in the presence of certain interaction effects (higher order terms).

- **Additivity, Non-Additivity, Saturation.** The effects model presented here, with (non-zero) interaction effects,  $(\alpha\beta)_{ij}$ , is often referred to as a **non-additive** model because we cannot reproduce the treatment means,  $\mu_{ij}$ , by *adding* main effects (and overall effect) alone; we also need to add  $(\alpha\beta)_{ij}$  to reproduce treatment means. (How do we get to non-additivity by one more addition?!...anyway...) We also say that the model is **saturated** in the sense that we cannot specify further (linear) mean model complexity/flexibility; we have reached the “saturation point” of (linear) mean model complexity by allowing each treatment level to have any value whatsoever. We might even consider the notion of model **over-saturation**: the case of having too much complexity in the sense of not being able to identify it,

i.e., not being able to identify parameter values, as illustrated by the effects model (with interaction) without constraints, discussed above. We may consider parsimony to be the flip-side of flexibility; we like to explain things as simply as possible. We might simplify our model (often supported by an F-test) by omitting, e.g., the interactions, to arrive at a simpler (**additive**) model,

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \epsilon_{ijk} \quad \epsilon_{ijk} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

$$i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n_{ij}.$$

(All quantities are as described previously in the non-additive/saturated model.) As we have discussed (Section 10.2), without interactions, we are left with a relatively simple interpretation in terms of main effects.

- **Strategy for Analysis.** As the above discussion may suggest, we often begin (if data permit) with a non-additive model, then test to see if we can reasonably omit interaction effects, which can simplify further analysis/interpretation. This process is discussed in [RS13, Sec. 13.5.1] and [KNNL05, Sec. 19.7]. See, in particular, [KNNL05, Fig. 19.11 pg. 848]. In some sense, this process is no more a matter of performing F v R F-tests, as we will continue to illustrate below. Thus, we are already in good stead.

### 10.6.1 E.g.: Factor Effects S2Zero Initial Analysis

We continue to use the seaweed grazing example, introduced above, to illustrate analysis in R using the factor effects model with sum-to-zero constraints.

If you haven't already recognized our approach, it may be helpful to realize that we are proceeding analogously to our one-way analysis of the factor effects model with sum-to-zero constraint, covered previously in §9.7.

```
> ## Default constraint/coding is treatment...:
>getOption("contrasts")

[1] "contr.treatment" "contr.treatment"

> ## ...unless factors have a different, overriding
> ## contrasts attribute...nope:
> sapply(case1301.df, attr, which="contrasts")

$Cover
NULL

$Block
NULL

$Treat
NULL

> ## Again, we can set constraints/coding globally...
> options(contrasts = rep("contr.sum",2))
>getOption("contrasts")

[1] "contr.sum" "contr.sum"

> ## ...or, we can assign (overriding) constraints/coding
> ## (perhaps not the same) to individual factors:
> contrasts(case1301.df$Block)<- contr.sum(levels(case1301.df$Block))
> contrasts(case1301.df$Treat)<- contr.sum(levels(case1301.df$Treat))
> sapply(case1301.df, attr, which="contrasts")

$Cover
NULL

$Block
  [,1] [,2] [,3] [,4] [,5] [,6] [,7]
B1     1     0     0     0     0     0     0
B2     0     1     0     0     0     0     0
B3     0     0     1     0     0     0     0
B4     0     0     0     1     0     0     0
```

```
B5    0    0    0    0    1    0    0
B6    0    0    0    0    0    1    0
B7    0    0    0    0    0    0    1
B8   -1   -1   -1   -1   -1   -1   -1

$Treat
 [,1] [,2] [,3] [,4] [,5]
C     1    0    0    0    0
L     0    1    0    0    0
Lf    0    0    1    0    0
Lff   0    0    0    1    0
f     0    0    0    0    1
ff    -1   -1   -1   -1   -1
```

Now, let's fit our first multi-way ANOVA model. We'll skip inspection of the (non-redundant)  $\mathbf{X}$  in R; we've looked at (some illustration of) it, above (perhaps in class), and have illustrated the `model.matrix` function before, so you should be able to reproduce it by hand or with R.

```
> case1301.lm<- lm(Cover ~ Block + Treat + Block:Treat,
+                      data=case1301.df)
> ## Estimated mean model coefficients (parameters) are given
> ## by the default printout, often of subsidiary interest
> ## in an ANOVA context:
> case1301.lm
```

Call:

```
lm(formula = Cover ~ Block + Treat + Block:Treat, data = case1301.df)
```

Coefficients:

|               | Block1 | Block2  | Block3  |
|---------------|--------|---------|---------|
| (Intercept)   | 28.625 | -19.125 | -15.375 |
| Block4        | 26.375 | -4.375  | 8.625   |
| Treat1        | 23.375 | -9.375  | -12.125 |
| Treat5        | 14.125 | -14.375 | -8.125  |
| Block4:Treat1 | 16.125 | -4.125  | 5.875   |
| Block1:Treat2 |        |         | -13.375 |
| Block2:Treat2 |        |         |         |
| Block3:Treat2 |        |         |         |
| Block4:Treat2 |        |         |         |

```

            3.875      3.625      11.125     -14.625
Block5:Treat2  Block6:Treat2  Block7:Treat2  Block1:Treat3
            7.125      -4.375      -6.625       6.625
Block2:Treat3  Block3:Treat3  Block4:Treat3  Block5:Treat3
            3.375      -9.125      -3.875      -5.125
Block6:Treat3  Block7:Treat3  Block1:Treat4  Block2:Treat4
            9.375      -2.875      12.875      11.625
Block3:Treat4  Block4:Treat4  Block5:Treat4  Block6:Treat4
           -12.875     -19.625      2.125      -2.875
Block7:Treat4  Block1:Treat5  Block2:Treat5  Block3:Treat5
            7.375      -6.125      -4.875       0.125
Block4:Treat5  Block5:Treat5  Block6:Treat5  Block7:Treat5
           16.875      -4.875      -5.875      10.875

> ## Typical regression summary, often of subsidiary interest
> ## in an ANOVA context:
> summary(case1301.lm)

```

Call:

```
lm(formula = Cover ~ Block + Treat + Block:Treat, data = case1301.df)
```

Residuals:

| Min     | 1Q     | Median | 3Q    | Max    |
|---------|--------|--------|-------|--------|
| -18.000 | -3.625 | 0.000  | 3.625 | 18.000 |

Coefficients:

|               | Estimate | Std. Error | t value | Pr(> t )     |
|---------------|----------|------------|---------|--------------|
| (Intercept)   | 28.6250  | 0.9648     | 29.670  | < 2e-16 ***  |
| Block1        | -19.1250 | 2.5525     | -7.493  | 1.30e-09 *** |
| Block2        | -15.3750 | 2.5525     | -6.023  | 2.31e-07 *** |
| Block3        | 12.6250  | 2.5525     | 4.946   | 9.68e-06 *** |
| Block4        | 26.3750  | 2.5525     | 10.333  | 8.58e-14 *** |
| Block5        | -4.3750  | 2.5525     | -1.714  | 0.09298 .    |
| Block6        | 8.6250   | 2.5525     | 3.379   | 0.00145 **   |
| Block7        | -5.6250  | 2.5525     | -2.204  | 0.03238 *    |
| Treat1        | 23.3750  | 2.1573     | 10.835  | 1.72e-14 *** |
| Treat2        | -9.3750  | 2.1573     | -4.346  | 7.17e-05 *** |
| Treat3        | -12.1250 | 2.1573     | -5.621  | 9.47e-07 *** |
| Treat4        | -20.8750 | 2.1573     | -9.677  | 7.35e-13 *** |
| Treat5        | 14.1250  | 2.1573     | 6.548   | 3.64e-08 *** |
| Block1:Treat1 | -14.3750 | 5.7076     | -2.519  | 0.01517 *    |
| Block2:Treat1 | -8.1250  | 5.7076     | -1.424  | 0.16105      |

```

Block3:Treat1  9.8750   5.7076   1.730  0.09003 .
Block4:Treat1 16.1250   5.7076   2.825  0.00687 **
Block5:Treat1 -4.1250   5.7076  -0.723  0.47336
Block6:Treat1  5.8750   5.7076   1.029  0.30849
Block7:Treat1 -13.3750  5.7076  -2.343  0.02330 *
Block1:Treat2  3.8750   5.7076   0.679  0.50045
Block2:Treat2  3.6250   5.7076   0.635  0.52837
Block3:Treat2 11.1250   5.7076   1.949  0.05713 .
Block4:Treat2 -14.6250  5.7076  -2.562  0.01359 *
Block5:Treat2  7.1250   5.7076   1.248  0.21797
Block6:Treat2 -4.3750   5.7076  -0.767  0.44712
Block7:Treat2 -6.6250   5.7076  -1.161  0.25149
Block1:Treat3  6.6250   5.7076   1.161  0.25149
Block2:Treat3  3.3750   5.7076   0.591  0.55708
Block3:Treat3 -9.1250  5.7076  -1.599  0.11644
Block4:Treat3 -3.8750  5.7076  -0.679  0.50045
Block5:Treat3 -5.1250  5.7076  -0.898  0.37371
Block6:Treat3  9.3750   5.7076   1.643  0.10701
Block7:Treat3 -2.8750  5.7076  -0.504  0.61677
Block1:Treat4 12.8750   5.7076   2.256  0.02868 *
Block2:Treat4 11.6250   5.7076   2.037  0.04721 *
Block3:Treat4 -12.8750  5.7076  -2.256  0.02868 *
Block4:Treat4 -19.6250  5.7076  -3.438  0.00122 **
Block5:Treat4  2.1250   5.7076   0.372  0.71130
Block6:Treat4 -2.8750  5.7076  -0.504  0.61677
Block7:Treat4  7.3750   5.7076   1.292  0.20250
Block1:Treat5 -6.1250  5.7076  -1.073  0.28858
Block2:Treat5 -4.8750  5.7076  -0.854  0.39728
Block3:Treat5  0.1250  5.7076   0.022  0.98262
Block4:Treat5 16.8750   5.7076   2.957  0.00481 **
Block5:Treat5 -4.8750  5.7076  -0.854  0.39728
Block6:Treat5 -5.8750  5.7076  -1.029  0.30849
Block7:Treat5 10.8750   5.7076   1.905  0.06274 .

---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 9.453 on 48 degrees of freedom  
 Multiple R-squared: 0.9192, Adjusted R-squared: 0.84  
 F-statistic: 11.61 on 47 and 48 DF, p-value: 1.119e-14

> ## Typical ANOVA table, often receives primary interest  
 > ## in an ANOVA context.

```
> anova(case1301.lm)

Analysis of Variance Table

Response: Cover
            Df  Sum Sq Mean Sq F value    Pr(>F)
Block        7 19105.5  2729.4 30.5454 1.296e-15 ***
Treat        5 23045.5  4609.1 51.5824 < 2.2e-16 ***
Block:Treat 35  6612.5   188.9  2.1144  0.008128 **
Residuals   48  4289.0    89.4
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

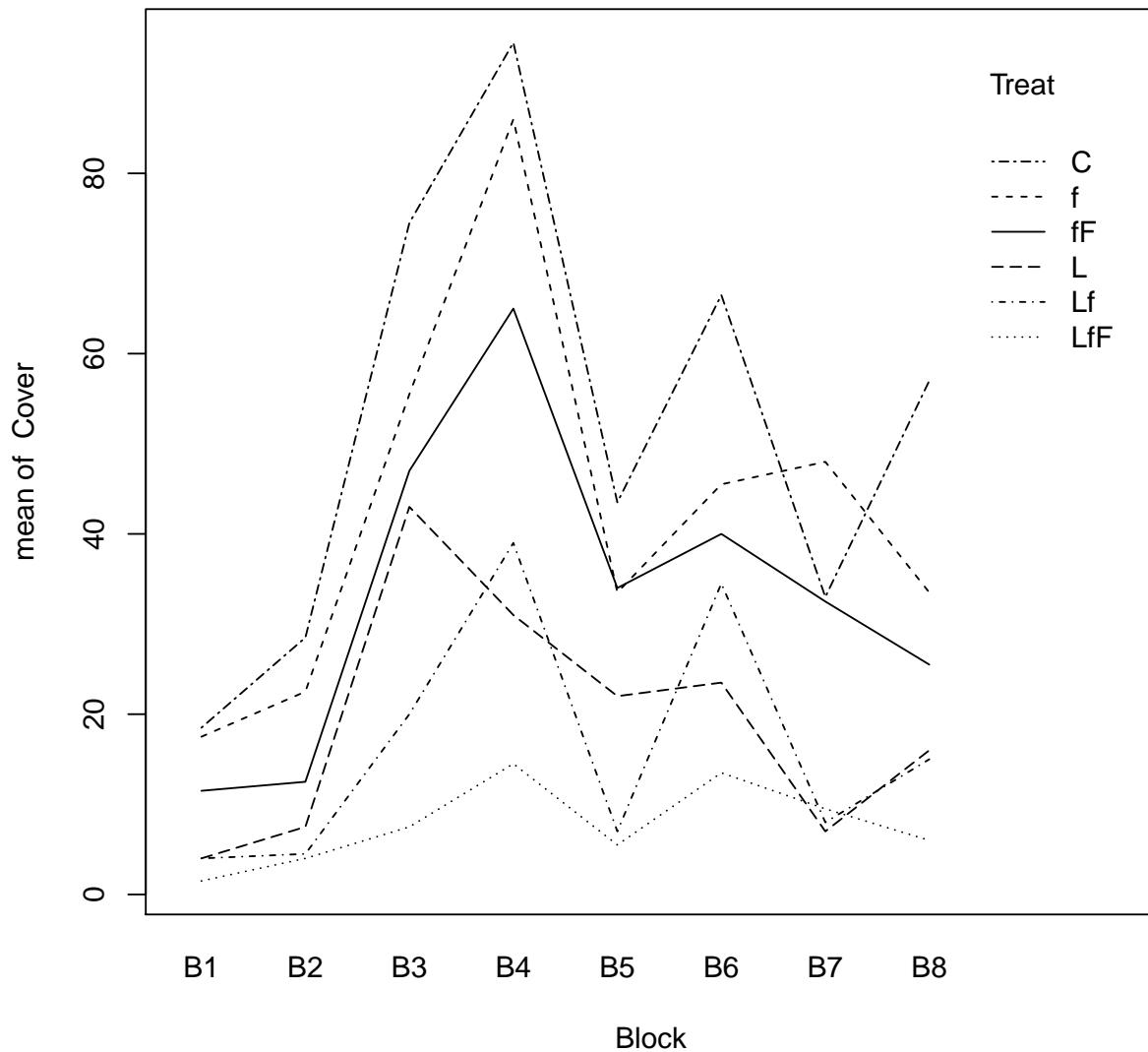
- What does the  $\mathbf{R} \mathbf{X}$  matrix look like? (Not in output...too big; we should have a fair idea of what it looks like at this point, given previous discussion.)
- What are the LS estimators/estimates of the parameters?
- How do we interpret these (estimated) parameters? Again, we may already have a fair understanding of the parameters/estimated parameters for the sum-to-zero constrained effects model. This understanding may help when making detailed inference, beyond a standard ANOVA table, i.e., for general  $\mathbf{C}\boldsymbol{\beta}$ .
- How do these estimates compare to those for the cell means model? We did not perform a cell-means model analysis for this two-way case, so we did not directly discuss  $\hat{\mu}_{ij}$ . Still, it's a fair question for someone who would brandish an ANOVA.
- What is the estimator/estimate of the variance,  $\sigma^2$  (or of the standard deviation,  $\sigma$ )? How would this compare with the estimate given by the cell means analysis if we were to have performed one?
- What are the estimated standard errors of the estimators of the parameters? (Think in terms of diagonal elements of a certain matrix, as we have discussed before.)

- R gives default t-tests for each of the parameters assuming a null value of zero by default. How are these tests computed? Are these tests interesting?
- How are the p-values for the above tests computed?
- What are the remaining quantities in the output of the `summary` function?

We might conclude that there are significant interaction effects as indicated by the default F-test given in the `anova` output, which would dictate the course of further analysis, as discuss above. Perhaps, we are being a bit hasty. Let's **diagnose** our model first before we proceed. (We follow [RS13, Sec. 13.3].)

Below, we construct an **interaction plot** based on averages of `Cover` at each treatment (not `Treat!`) level (i.e.,  $\bar{Y}_{ij\cdot}$ , which, incidentally, are the least squares fitted values,  $\hat{Y}_{ijk}$  (either  $k = 1$  or  $k = 2$  of course!) for the cell means model or the saturated (non-additive) effects model. (Recall that our non-additive model is equivalent to the cell means model so that treatment (cell) averages minimize the LS criterion.). In other words, we plot fitted values versus our covariates (factors) to see if we can diagnose non-additivity or perhaps some departure from model assumptions, i.e., does there appear to be interaction or is there something else going on? You might want to revisit Section 10.2 before proceeding.

```
> ## Similar to Display 13.7 in R&S:
> with(case1301.df, {interaction.plot(x.factor=Block,
+                                         trace.factor=Treat,
+                                         response=Cover)})
```

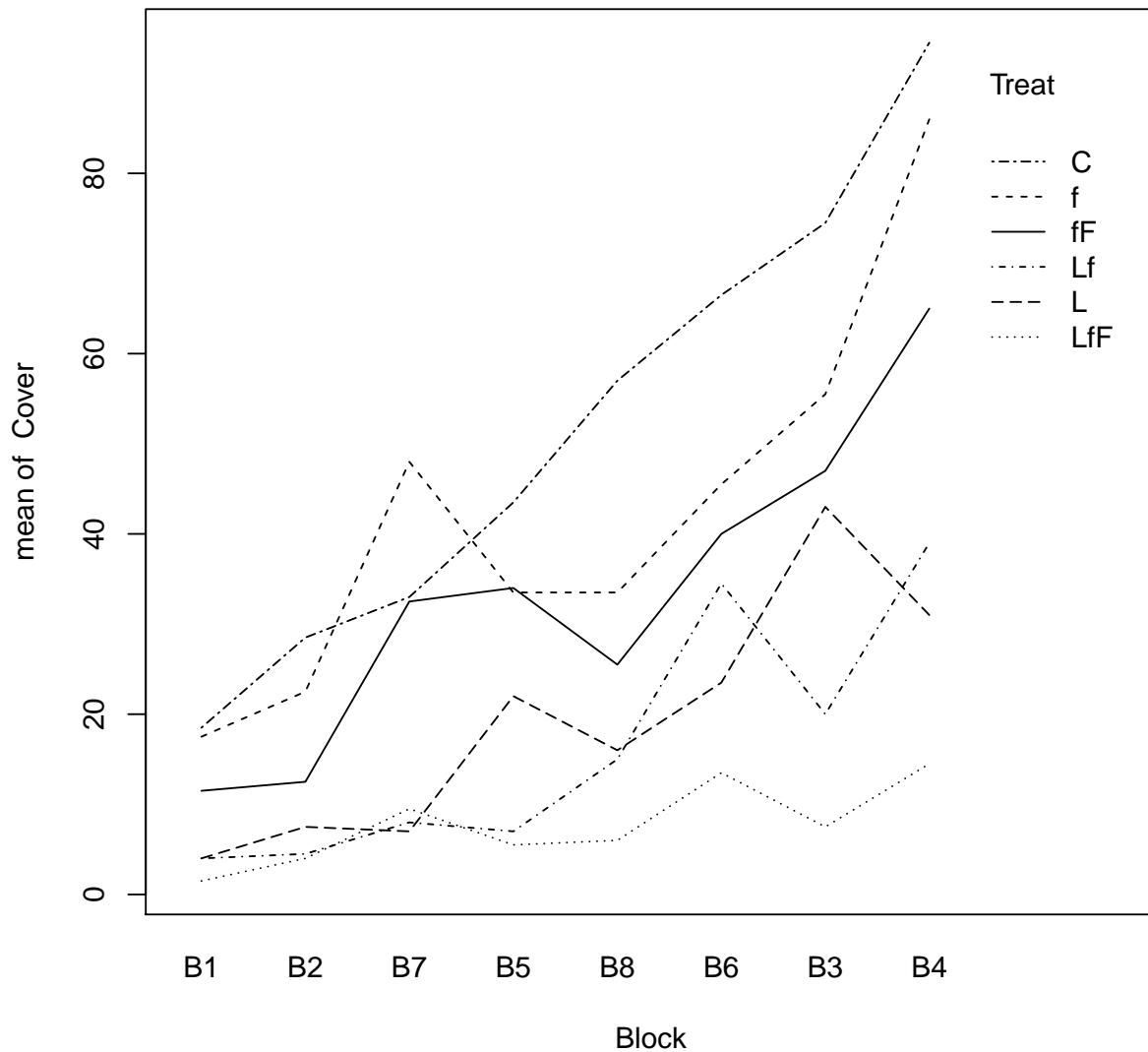


Let's edit the interaction plot to be more like [RS13, Display 13.7] (in terms of Block levels, at least).

```
> ## Perhaps we want to mimic Display 13.7 more closely:
> (ybidd<- tapply(case1301.df$Cover, case1301.df$Block, mean))

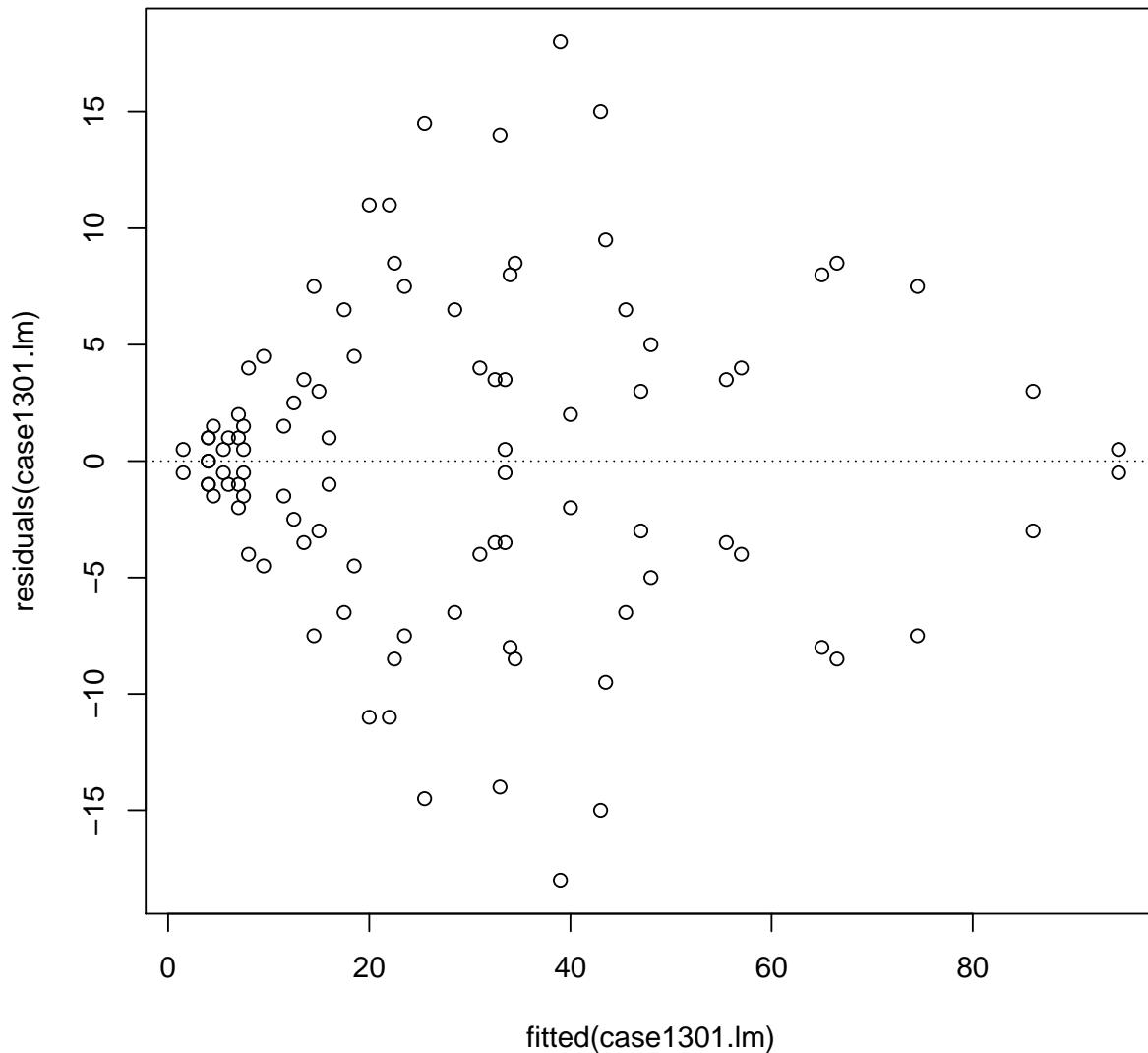
      B1      B2      B3      B4      B5      B6      B7      B8 
  9.50  13.25  41.25  55.00  24.25  37.25  23.00  25.50
```

```
> (indx<- order(ybidd))  
[1] 1 2 7 5 8 6 3 4  
  
> levels(case1301.df$Block)  
[1] "B1" "B2" "B3" "B4" "B5" "B6" "B7" "B8"  
  
> levels(case1301.df$Block)[indx]  
[1] "B1" "B2" "B7" "B5" "B8" "B6" "B3" "B4"  
  
> tmp.df<- case1301.df  
> tmp.df$Block<- factor(case1301.df$Block,  
+                           levels=levels(case1301.df$Block)[indx])  
> with(tmp.df, {interaction.plot(x.factor=Block,  
+                                    trace.factor=Treat,  
+                                    response=Cover)})
```



I (and R&S) think that we may not be seeing interaction, but may be seeing non-constant variance, typically seen in proportion or percent data. The interaction plot may not be the best way to illustrate non-constant variance—it's using averages instead of observations—still, it does hint at non-constant variance. As we may recall, residuals are often a good way to diagnose non-constant variance ([RS13, Display 13.8]).

```
> ## Compare to R&S Display 13.8 (classic):  
> plot(residuals(case1301.lm) ~ fitted(case1301.lm))  
> abline(h=0, lty=3)
```



Also, theory suggests that proportions or percentages have non-constant variance. Further, the pattern in the residual plot above is entirely consistent with what the theory suggests—classic. Base on this theory, a typical

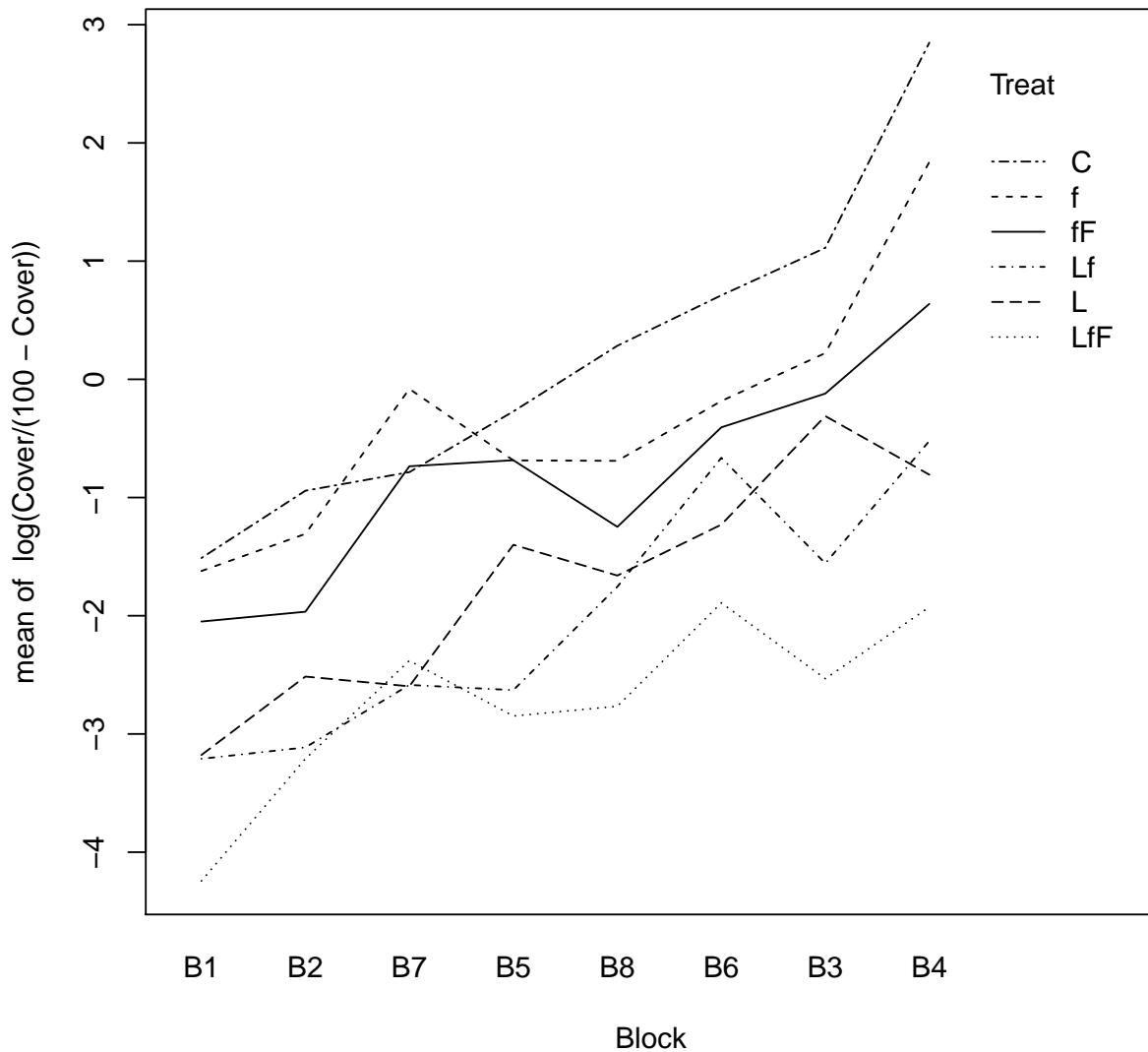
transformation of a proportion,  $Y$ , is the logit transformation:

$$\text{logit}(Y) = \log(Y/(1 - Y))$$

([KNNL05, Sec. 18.5]).

We may create a new variable in our data frame, or simply specify the transformation in the model formula, directly. (Recall that the response variable, `Cover`, is a percentage, not a proportion.) The interaction plot, based on the transformed response, suggests that the transformation stabilizes variance.

```
> ## Compare to R&S Display 13.9...better?
> with(tmp.df, {interaction.plot(x.factor=Block,
+                                   trace.factor=Treat,
+                                   response=log(Cover/(100-Cover))))}
```

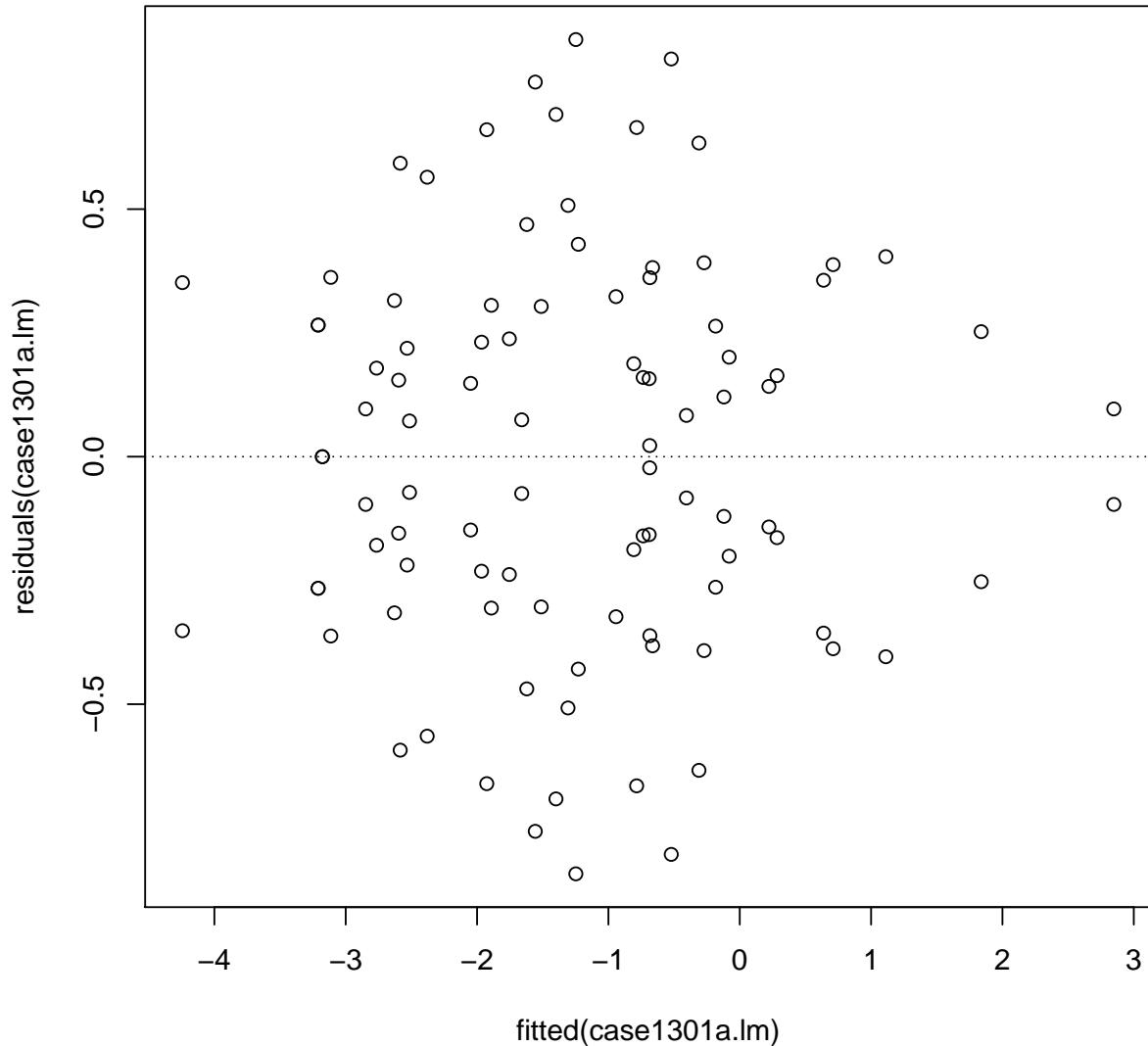


Again, residuals (next Chunk) may allow a better diagnosis. Your eye may suggest remaining heteroskedasticity, but I think the residual pattern may be attributed more to the concentration of values near mid-level proportions, and less to residual heteroskedasticity. In any case, I am not too concerned about the pattern.

Incidentally, why the peculiar-looking reflection of points about the horizontal zero line? If you turn your head sideways and let your eyes wander a bit, you can see an elderly lady or a young lady, depending on which way

you turn your head...just kidding. Relatedly (and slightly more seriously), this symmetry of residuals can be seen in the brief residual diagnostic of the `summary` function for the non-transformed case, above (and for the transformed case had we used `summary` in that case).

```
> case1301a.lm<- lm(log(Cover/(100-Cover)) ~ Block + Treat +
+                         Block:Treat,
+                         data=case1301.df)
> ## Improved
> plot(residuals(case1301a.lm) ~ fitted(case1301a.lm))
> abline(h=0, lty=3)
```



As we should know by now (see Section 10.2), additivity (no interaction) is exhibited (ideally!) by parallel lines in an interaction plot, indicating that mean differences among Factor A (B) levels are the same for each level of Factor B (A). In the last interaction plot, above, differences of averages (of transformed values) among the levels of **Treat** are not exactly the same across the levels of **Block**, but it seems reasonable to argue that the differences are comparable across the levels of **Block**, and we may suggest that the departure from parallel lines may simply be due to sampling error variation (recall that

we only have  $n_{ij} = 2$  observations to estimate the treatment (cell) means).

### 10.6.2 E.g.: Effects S2Zero ANOVA For Common $C\beta$

Following the above analysis, let's continue on the logit scale. The next chunk shows a standard ANOVA table, which verifies the ANOVA table shown in [RS13, Display 13.10].

```
> anova(case1301a.lm)

Analysis of Variance Table

Response: log(Cover/(100 - Cover))
            Df  Sum Sq Mean Sq F value Pr(>F)
Block        7  76.239 10.8912 35.9634 <2e-16 ***
Treat        5  96.993 19.3986 64.0553 <2e-16 ***
Block:Treat 35 15.230  0.4352  1.4369 0.1209
Residuals   48 14.536  0.3028
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We skip the estimated parameter printouts, which we covered before the transformation. Interpretation is the same, except it's now in terms of logit(proportion cover). Besides, as I should have said earlier, the parameter estimates in ANOVA are typically of subsidiary interest. Similarly, we skip  $\mathbf{X}$  and  $\boldsymbol{\beta}$  (or estimate thereof), as we have discussed these previously.
- The ANOVA table simply summarizes typical tests of the significance of factor variables or their interactions. It is often featured in a course like this. Fine. But, we don't need it, because we have our F vs. R and linear combinations approaches to infer more generally about linear combination of parameters. Still, the tests are likely to be of use, so we should discuss the table, at least for historical reasons and because ANOVA tables are featured in textbooks and

computer printouts. (Admittedly, it may be relatively convenient as the numbers of factors/interactions increase.)

Again, we ask, what is ‘ANalysis of VAriance (ANOVA)?’ In the one-way case, the answer was relatively simple and uninteresting (Section 9.10). Here, the answer is a mildly more complicated and interesting. We skip some details, which typically accompany a discussion of “hand computations”. Though an argument may be made that discussion of these computations may facilitate conceptual understanding, ultimately, I believe they serve little purpose and in some sense seem out of date to me. See [KNNL05, Sec. 19.4] for more details behind balanced ANOVA hand computations.

As in the one-way case (Section 9.10), we have a decomposition of sum-of-squares,

$$\text{SSTO} = \text{SSTR} + \text{SSE}.$$

Now, in the **BALANCED** two-way case, we can further decompose SSTR into components associated with the factor variables and their interaction(s):

$$\text{SSTR} = \text{SSA} + \text{SSB} + \text{SSAB},$$

where

- **SSA** is “factor A sum-of-squares,” a measure of the response variability that is associated with Factor A. For our running example (see the latest output from the `anova` function), this is 76.239 (treating `Block` as A and `Treat` as B).
- **SSB** is “factor B sum-of-squares,” a measure of the response variability that is associated with Factor A. 96.993.

- **SSAB** is “interaction sum-of-squares,” a measure of the response variability that is associated with the interaction between Factors A and B. 15.230.
- [RS13, Display 13.10] shows SSTR as “Between groups” sum-of-squares: 188.4622, which is, in this **BALANCED** case, the sum of the above SS. Computer printouts may or may not show SSTR (Between groups SS).

And, as in the one-way case (Section 9.10), we also have a corresponding decomposition of the total degrees of freedom:

$$(n_T - 1) = (ab - 1) + (n_T - ab),$$

where, analogously to the decomposition of SSTR, we can further decompose treatment degrees of freedom,  $(ab - 1)$ , as

$$(ab - 1) = (a - 1) + (b - 1) + (a - 1)(b - 1),$$

where

- $(n_T - 1)$  is the “**total degrees of freedom**,” as in the one-way case; 95 in the running example, though not shown directly by `anova`.
- $(ab - 1)$  (47) is the “**treatment degrees of freedom**” (we have  $ab = (8)(6) = 48$  treatment levels, right?)
- $(n_T - ab)$  is the “**error degrees of freedom**” (or “residual degrees of freedom”); 48.
- $(a - 1)$  is “**factor A degrees of freedom**”; 7.
- $(b - 1)$  is “**factor B degrees of freedom**”; 5.
- $(a - 1)(b - 1)$  “**AB interaction degrees of freedom**”; 35.

- Note that the degrees of freedom associated with A, B, and AB, are merely the number of columns associated with each factor or interaction in the (non-redundant)  $\mathbf{X}$  matrix.

The sum-of-squares and degrees of freedom are used to construct further entries in the standard ANOVA table.

- As in the one-way case, we divide sums-of-squares by their respective degrees of freedom to get mean squares, MSTO, MSTR and MSE, although MSTO is not typically used (it's just the sample variance of the  $Y$  values), and these mean squares are *not* additive like the above SS's and df's. Again, [RS13, Display 13.10] shows MSTR as "Between group" mean square and MSE as "Within group" mean square.
- In addition, in the two-way case, we have MSA, MSB, and MSAB (mean square for `Block`, mean square for `Treat` and mean square for `Block:Treat` interaction in the running example). Look at [RS13, Display 13.10] or the last `anova` output for values. Again, computer programs may not show SSTR or MSTR and usually do not show SSTO or MSTO.
- Also, we obtain various F-statistics and associated p-values, computed under a null distribution, which we will discuss in class. Note that the presentation of the ratio of mean squares as F-statistics in an ANOVA table seems somehow "classic" (dated?) to me. (Granted, the tests are often of interest.) We can (have/will) use the linear combinations ( $\mathbf{C}\boldsymbol{\beta}$ ) approach (or F v R approach) to get the same F-statistics. See, once again, §6.7. In other words, the ANOVA table may be convenient, but is somehow extra, automated output that is not necessary given the alternative approaches that we've discussed (F v R or linear combinations). Still, with more complicated, higher-way layouts, standard ANOVA tables may be convenient.

- Note, ANOVA tables may be ambiguous about what is being tested, in either the **BALANCED** or **UNBALANCED** cases; see Section 10.8, below, for more about this.

### 10.6.3 E.g.: Factor Effects S2Zero F v R & $C\beta$ Approach

Recall, we want to say something about the effects of grazing on seaweed regeneration, and we know that absence of interaction, as indicated, above, will tend to simplify our inferences for the **Treat** factor. And, we are not particularly interested in the effects of location (i.e., blocks, i.e., **Block** factor levels). But, before omitting interactions and proceeding to infer about grazing effects, let's, once again, illustrate the use of our F v R (aka extra sum of squares) and linear combinations ( $C\beta$ ) approaches as **alternatives to using the automatic F-tests of the ANOVA table**, illustrated in the previous section.

In particular, we look at the interaction term. Of course, we already know, from the ANOVA table and our previous discussion, that there does not appear to be serious, significant interaction, on the logit scale. But, again, it doesn't hurt to continue building our familiarity with these alternative, more general approaches to inference. We will use these later, when there is not an automatically generated ANOVA table that happens to summarize our interesting inferences for us.

Continuing in R, first with the **F v R approach**, we could use `lm`, now omitting `Block:Treat` from the model formula, to produce a reduced model. Instead, we use the `update` function. Again, that some sum-of-squares, etc., do not change between F and R models generally only happens in **BALANCED** cases. Do you begin to see a potential ambiguity of tests as presented in ANOVA tables? See §10.8, below.

```
> ## Reduced model, without interaction term:
> case1301aR.lm<- update(case1301a.lm,. ~ . - Block:Treat)
> ## Just to illustrate effect of balance, otherwise...
> ## ...NO...this is a sequential ANOVA table...
> anova(case1301aR.lm)
```

```
Analysis of Variance Table
```

```
Response: log(Cover/(100 - Cover))
          Df Sum Sq Mean Sq F value    Pr(>F)
Block      7 76.239 10.8912 30.368 < 2.2e-16 ***
Treat      5 96.993 19.3986 54.090 < 2.2e-16 ***
Residuals 83 29.767  0.3586
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> ##...just go to F v R test:
> anova(case1301aR.lm, case1301a.lm)
```

```
Analysis of Variance Table
```

```
Model 1: log(Cover/(100 - Cover)) ~ Block + Treat
Model 2: log(Cover/(100 - Cover)) ~ Block + Treat + Block:Treat
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     83 29.767
2     48 14.536 35     15.23 1.4369 0.1209

> ## which gives the same test for interaction as given
> ## by anova(case1301a.lm), as promised.
```

The **linear combinations approach**, using `glh.test`, at least, is a bit less convenient because it requires us to recall the non-redundant (coded or “regression”) form of  $\mathbf{X}$ . More particularly, it requires us to interpret the parameters in the associated  $\boldsymbol{\beta}$  vector—after all, we want to infer about particular linear combinations of parameters, so we should know what  $\boldsymbol{\beta}$  is!

Which elements of  $\boldsymbol{\beta}$  associated with the interaction, i.e., with the  $(\alpha\beta)_{ij}$ , i.e., with the `Treat:Block` model formula term?

```
> ## I build C matrix in stages (not necessary, but seems convenient).
> ##
> a<- 8; b<- 6 ## number of factor levels
> Ca<- matrix(0,nrow=(a-1)*(b-1), ncol=a-1)
> Cb<- matrix(0,nrow=(a-1)*(b-1), ncol=b-1)
> Cab<- diag((a-1)*(b-1))
> Cmat<- cbind(0,Ca,Cb,Cab) ## C matrix
> d<- rep(0,(a-1)*(b-1)) ## null CBeta value
```

```
> library(gmodels)
> glh.test(reg=case1301a.lm, cm=Cmat, d=d)

Test of General Linear Hypothesis
Call:
glh.test(reg = case1301a.lm, cm = Cmat, d = d)
F = 1.4369, df1 = 35, df2 = 48, p-value = 0.1209

> ## Result is same as F v R approach or anova(case1301a.lm)
> ## of course!
```

#### 10.6.4 E.g.: Effects 2Zero Summary

Thus, we've seen how traditional ANOVA tables are used to present tests for certain, commonly interesting, linear combinations of parameters, which we can also infer about using our usual F v R and  $\mathbf{C}\boldsymbol{\beta}$  approaches. However we approached these inferences, results suggest that we feel comfortable proceeding to infer about grazing effects with the additive model, i.e., without the interaction term. But, before we do that, we mimic our one-way presentation and cover treatment constraints/coding. While, perhaps, not as popular as effects sum-to-zero constraints/coding, treatment constraints/coding is the default coding in R. I think it's worth knowing about both, obviously.

### 10.7 Factor Effects: Treatment Constraints/Coding

We essentially repeat the above sum-to-zero analyses, now in terms of treatment constraints/coding as an alternative way to resolve redundancy among the columns of  $\mathbf{X}$ . In particular, we set particular parameters to zero:

$$\begin{aligned}\alpha_1 &= 0 \\ \beta_1 &= 0 \\ (\alpha\beta)_{1j} &= 0 \quad j=1, \dots, b, \\ (\alpha\beta)_{i1} &= 0 \quad i=1, \dots, a.\end{aligned}$$

In other words, similar to the treatment constraint/coding in the one-way case, we work with the remaining  $\alpha_i$ ,  $\beta_j$ ,  $(\alpha\beta)_{ij}$   $i = 2, \dots, a$   $j = 2, \dots, b$ . Note that R sets the first effect parameter to zero by default; SAS sets the last effect parameter to zero when using treatment coding.

- What does the non-redundant  $\boldsymbol{\beta}$  look like? (more in class)
- What does the non-redundant  $\mathbf{X}$  look like? (more in class) (BTW, all interaction term (re-coded) columns can be obtained by multiplying (element-wise!) each (re-coded) column associated with one factor with each (re-coded) column associated with another factor, just as with the sum-to-zero constraints/coding.)
- Again, we are simply performing regression with specially coded “X” variables, one for each column in  $\mathbf{X}$ , and with associated coefficients having a particular interpretations. Incidentally, what is the “reference cell”? (Recall treatment coding is aka cell reference coding.)
- We’ll see how to implement this coding/constraints in R, shortly.
- Higher way?
- Non-additive model?

Notice (if we have not already) the interpretation of parameters in this factor effects model with treatment (cell reference) constraints/coding:

- As with sum-to-zero constraints/coding, we can relate the cell means and effects as

$$\mu_{ij} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij}.$$

Indeed, this (non-additive or saturated...more shortly) effects model is equivalent to the cell means model and to the effects model with sum-to-zero constraints/coding: their respective  $\mathbf{X}$  matrices span the same covariate/regressor space, each gives the same predicted values, same residuals, same estimate of  $\sigma^2$ , same overall F-test...perhaps more on this point later.

- **Overall Effect.** If we look at  $i = 1$  and  $j = 1$ , we see that

$$\mu_{11} = \mu_{..},$$

so that, now, with treatment constraints,  $\mu_{..}$  may still, of course, be considered an overall effect common to all observations, but we see now it is the mean of the **reference cell or reference treatment** defined by the first level of factor A and the first level of Factor B.

- **Main Effects of Factor A.** With treatment constraints, for  $j = 1$ , we see

$$\mu_{i1} = \mu_{..} + \alpha_i,$$

or

$$\alpha_i = \mu_{i1} - \mu_{..},$$

or

$$\alpha_i = \mu_{i1} - \mu_{11},$$

so that,  $\alpha_i$  may still be interpreted as the deviation from the overall effect (not the overall mean now!) due to being in the  $i$ th level of factor A, but, more particularly, it is the deviation from the mean of the reference cell due to being in the  $i$ th level of Factor A (because we're

not changing the reference level of B,  $j = 1$ , here). Again, inference for main effects may not be sensible in the presence of interaction.

- **Main Effects of Factor B.** With treatment constraints, for  $i = 1$ , we see

$$\mu_{1j} = \mu_{..} + \beta_j,$$

or

$$\beta_j = \mu_{1j} - \mu_{..},$$

or

$$\beta_j = \mu_{1j} - \mu_{11},$$

so that,  $\beta_j$  may still be interpreted as the deviation from the overall effect (not the overall mean now!) due to being in the  $j$ th level of factor B, but, more particularly, it is the deviation from the mean of the reference cell due to being in the  $j$ th level of Factor B (because we're not changing the reference level of A,  $i = 1$ , here). Again, inference for main effects may not be sensible in the presence of interaction.

- **Interaction Effects.** For  $i, j > 1$ , we have

$$\begin{aligned} (\alpha\beta)_{ij} &= \mu_{ij} - (\mu_{..} + \alpha_i + \beta_j) \\ &= \mu_{ij} - (\mu_{11} + \alpha_i + \beta_j) \end{aligned}$$

so that  $(\alpha\beta)_{ij}$ , again, is the additional effect, beyond the main effects, necessary to reproduce the  $ij$ th mean, now starting from the reference cell mean, not from the overall mean. Similar to the interpretation for interaction effects in the sum-to-zero constraints/coding case, if we hold, e.g.,  $j$  constant, we see that we may interpret  $(\alpha\beta)_{ij}$  as an additional effect, beyond main effects, of the  $i$ th level of A, due to being at level  $j$  of factor B. (Or, If we hold, e.g.,  $i$  constant, we see that we may interpret  $(\alpha\beta)_{ij}$  as an additional effect of the  $j$ th level of B, due to being at level  $i$  of factor A.) In other words, the effects of factor A (B) depends on the level of factor B (A), and we say that

the factors **interact** and call  $(\alpha\beta)_{ij}$  an **interaction effect**. Again, we've discussed interaction (Section 10.2), and we have warned about the potential nonsensicality of inferring main effects in the presence of certain interaction effects.

- **Additivity, Non-Additivity, Saturation.** The same discussion as we gave with the sum-to-zero effects model applies equally here: The effects model presented here, with (non-zero) interaction effects,  $(\alpha\beta)_{ij}$ , is often referred to as a **non-additive** model because we cannot reproduce the treatment means,  $\mu_{ij}$ , by *adding* main effects (and overall effect...now not the overall mean but the reference cell mean) alone; we also need  $(\alpha\beta)_{ij}$  to reproduce treatment means. We also say that the model is **saturated** in the sense that we cannot specify further (linear) mean model complexity/flexibility; we have reached the “saturation point” of (linear) mean model complexity by allowing each treatment level to have any value whatsoever. We might even consider the notion of model **over-saturation**: the case of having too much complexity in the sense of not being able to identify it, i.e., not being able to identify parameter values, as illustrated by the effects model (with interaction) without constraints, discussed above. Again, we may consider parsimony to be the flip-side of flexibility; we like to explain things as simply as possible. We might simplify our model by omitting, e.g., the interactions, to arrive at a simpler (**additive**) model,

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \epsilon_{ijk} \quad \epsilon_{ijk} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

$$i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n_{ij}.$$

All quantities are as described previously in the non-additive/saturated model. As we have discussed (Section 10.2),

without interactions, we are left with a relatively simple interpretation in terms of main effects. (NOTE: In an attempt to avoid unnecessary confusion, our discussion of ANOVA components in Section 10.2 used the sum-to-zero interpretation of effects, to allow the somehow more intuitive interpretation of the overall effect as the overall mean, not the reference cell mean.)

- **Strategy for Analysis.** Same as given with the sum-to-zero effects model: As the above discussion may suggest, we often begin (if data permit) with a non-additive model, then test to see if we can reasonably omit interaction effects, which can simplify further analysis/interpretation. This process is discussed in [RS13, Sec. 13.5.1] and [KNNL05, Sec. 19.7]. See, in particular, [KNNL05, Fig. 19.11 pg. 848]. In some sense, this process is no more a matter of performing F v R F-tests, as we will continue to illustrate below. Thus, we are already in good stead.

### 10.7.1 E.g.: Factor Effects Trmt Initial Analysis

As mentioned before, treatment constraints are the default in R for (unordered) factors, but, we've been tinkering with such matters, so, first, we ensure treatment constraints before proceeding.

```
> ## Default constraint/coding is treatment, but may have changed:
>getOption("contrasts")
[1] "contr.sum" "contr.sum"

> ## Do factors have an overriding contrasts attribute?:
> sapply(case1301.df, attr, which="contrasts")

$Cover
NULL
```

```

$Block
[,1] [,2] [,3] [,4] [,5] [,6] [,7]
B1    1     0     0     0     0     0     0
B2    0     1     0     0     0     0     0
B3    0     0     1     0     0     0     0
B4    0     0     0     1     0     0     0
B5    0     0     0     0     1     0     0
B6    0     0     0     0     0     1     0
B7    0     0     0     0     0     0     1
B8   -1    -1    -1    -1    -1    -1    -1

$Treat
[,1] [,2] [,3] [,4] [,5]
C     1     0     0     0     0
L     0     1     0     0     0
Lf    0     0     1     0     0
LFF   0     0     0     1     0
f     0     0     0     0     1
ff    -1    -1    -1    -1    -1

> ## Again, we can set constraints/coding globally...
> options(contrasts = rep("contr.treatment",2))
>getOption("contrasts")

[1] "contr.treatment" "contr.treatment"

> ## ...or, we can assign (overriding) constraints/coding
> ## (perhaps not the same) to individual factors:
> contrasts(case1301.df$Block)<- contr.treatment(levels(case1301.df$Block))
> contrasts(case1301.df$Treat)<- contr.treatment(levels(case1301.df$Treat))
> sapply(case1301.df, attr, which="contrasts")

$Cover
NULL

$Block
B2 B3 B4 B5 B6 B7 B8
B1 0 0 0 0 0 0 0
B2 1 0 0 0 0 0 0
B3 0 1 0 0 0 0 0
B4 0 0 1 0 0 0 0
B5 0 0 0 1 0 0 0

```

```
B6  0  0  0  0  1  0  0
B7  0  0  0  0  0  1  0
B8  0  0  0  0  0  0  1
```

```
$Treat
  L Lf LfF f fF
C  0  0  0  0  0
L  1  0  0  0  0
Lf  0  1  0  0  0
LfF 0  0  1  0  0
f   0  0  0  1  0
fF  0  0  0  0  1
```

Now, let's repeat the fit of our first multi-way ANOVA model—using the logit transform—now using treatment constraints. We'll skip inspection of the (non-redundant)  $\mathbf{X}$  in R; we've looked at it, above, and have illustrated the `model.matrix` function before, so you should be able to reproduce it by hand or with R.

Whether using sum-to-zero constraints or treatment constraints, nothing changes for the R model formula in our seaweed grazing example. See Section 10.5, above, for discussion of R model formulae.

```
> case1301TC.lm<- lm(log(Cover/(100-Cover)) ~ Block + Treat +
+                         Block:Treat,
+                         data=case1301.df)
> ## Estimated mean model coefficients (parameters) are given
> ## by the default printout, often of subsidiary interest
> ## in an ANOVA context:
> case1301TC.lm
```

Call:

```
lm(formula = log(Cover/(100 - Cover)) ~ Block + Treat + Block:Treat,
   data = case1301.df)
```

Coefficients:

|             |         |         |
|-------------|---------|---------|
| (Intercept) | BlockB2 | BlockB3 |
| -1.51180    | 0.56945 | 2.62407 |
| BlockB4     | BlockB5 | BlockB6 |
| 4.35979     | 1.24023 | 2.22249 |
| BlockB7     | BlockB8 | TreatL  |

|                  |                  |                  |
|------------------|------------------|------------------|
| 0.72672          | 1.79553          | -1.66625         |
| TreatLf          | TreatLfF         | Treatf           |
| -1.69847         | -2.73167         | -0.10991         |
| TreatffF         | BlockB2:TreatL   | BlockB3:TreatL   |
| -0.53729         | 0.09409          | 0.24314          |
| BlockB4:TreatL   | BlockB5:TreatL   | BlockB6:TreatL   |
| -1.98856         | 0.53837          | -0.27361         |
| BlockB7:TreatL   | BlockB8:TreatL   | BlockB2:TreatLf  |
| -0.14561         | -0.27759         | -0.47300         |
| BlockB3:TreatLf  | BlockB4:TreatLf  | BlockB5:TreatLf  |
| -0.97068         | -1.67106         | -0.65899         |
| BlockB6:TreatLf  | BlockB7:TreatLf  | BlockB8:TreatLf  |
| 0.32387          | -0.10170         | -0.33965         |
| BlockB2:TreatLfF | BlockB3:TreatLfF | BlockB4:TreatLfF |
| 0.46375          | -0.91318         | -2.04250         |
| BlockB5:TreatLfF | BlockB6:TreatLfF | BlockB7:TreatLfF |
| 0.15526          | 0.12955          | 1.13688          |
| BlockB8:TreatLfF | BlockB2:Treatf   | BlockB3:Treatf   |
| -0.31762         | -0.25544         | -0.78035         |
| BlockB4:Treatf   | BlockB5:Treatf   | BlockB6:Treatf   |
| -0.89989         | -0.30425         | -0.78442         |
| BlockB7:Treatf   | BlockB8:Treatf   | BlockB2:TreatffF |
| 0.81413          | -0.86358         | -0.48627         |
| BlockB3:TreatffF | BlockB4:TreatffF | BlockB5:TreatffF |
| -0.69556         | -1.67246         | 0.12450          |
| BlockB6:TreatffF | BlockB7:TreatffF | BlockB8:TreatffF |
| -0.57956         | 0.58699          | -0.99454         |

```
> ## Typical regression summary, often of subsidiary interest
> ## in an ANOVA context:
> summary(case1301TC.lm)
```

Call:

```
lm(formula = log(Cover/(100 - Cover)) ~ Block + Treat + Block:Treat,
   data = case1301.df)
```

Residuals:

| Min     | 1Q      | Median | 3Q     | Max    |
|---------|---------|--------|--------|--------|
| -0.8426 | -0.2752 | 0.0000 | 0.2752 | 0.8426 |

Coefficients:

| Estimate | Std. Error | t value | Pr(> t ) |
|----------|------------|---------|----------|
|----------|------------|---------|----------|

|                  |          |         |        |          |     |
|------------------|----------|---------|--------|----------|-----|
| (Intercept)      | -1.51180 | 0.38913 | -3.885 | 0.000313 | *** |
| BlockB2          | 0.56945  | 0.55031 | 1.035  | 0.305960 |     |
| BlockB3          | 2.62407  | 0.55031 | 4.768  | 1.77e-05 | *** |
| BlockB4          | 4.35979  | 0.55031 | 7.922  | 2.89e-10 | *** |
| BlockB5          | 1.24023  | 0.55031 | 2.254  | 0.028820 | *   |
| BlockB6          | 2.22249  | 0.55031 | 4.039  | 0.000193 | *** |
| BlockB7          | 0.72672  | 0.55031 | 1.321  | 0.192907 |     |
| BlockB8          | 1.79553  | 0.55031 | 3.263  | 0.002036 | **  |
| TreatL           | -1.66625 | 0.55031 | -3.028 | 0.003955 | **  |
| TreatLf          | -1.69847 | 0.55031 | -3.086 | 0.003360 | **  |
| TreatLff         | -2.73167 | 0.55031 | -4.964 | 9.12e-06 | *** |
| Treatf           | -0.10991 | 0.55031 | -0.200 | 0.842541 |     |
| Treatff          | -0.53729 | 0.55031 | -0.976 | 0.333791 |     |
| BlockB2:TreatL   | 0.09409  | 0.77826 | 0.121  | 0.904278 |     |
| BlockB3:TreatL   | 0.24314  | 0.77826 | 0.312  | 0.756075 |     |
| BlockB4:TreatL   | -1.98856 | 0.77826 | -2.555 | 0.013839 | *   |
| BlockB5:TreatL   | 0.53837  | 0.77826 | 0.692  | 0.492423 |     |
| BlockB6:TreatL   | -0.27361 | 0.77826 | -0.352 | 0.726698 |     |
| BlockB7:TreatL   | -0.14561 | 0.77826 | -0.187 | 0.852372 |     |
| BlockB8:TreatL   | -0.27759 | 0.77826 | -0.357 | 0.722894 |     |
| BlockB2:TreatLf  | -0.47300 | 0.77826 | -0.608 | 0.546210 |     |
| BlockB3:TreatLf  | -0.97068 | 0.77826 | -1.247 | 0.218360 |     |
| BlockB4:TreatLf  | -1.67106 | 0.77826 | -2.147 | 0.036859 | *   |
| BlockB5:TreatLf  | -0.65899 | 0.77826 | -0.847 | 0.401333 |     |
| BlockB6:TreatLf  | 0.32387  | 0.77826 | 0.416  | 0.679161 |     |
| BlockB7:TreatLf  | -0.10170 | 0.77826 | -0.131 | 0.896581 |     |
| BlockB8:TreatLf  | -0.33965 | 0.77826 | -0.436 | 0.664487 |     |
| BlockB2:TreatLff | 0.46375  | 0.77826 | 0.596  | 0.554050 |     |
| BlockB3:TreatLff | -0.91318 | 0.77826 | -1.173 | 0.246438 |     |
| BlockB4:TreatLff | -2.04250 | 0.77826 | -2.624 | 0.011605 | *   |
| BlockB5:TreatLff | 0.15526  | 0.77826 | 0.199  | 0.842719 |     |
| BlockB6:TreatLff | 0.12955  | 0.77826 | 0.166  | 0.868492 |     |
| BlockB7:TreatLff | 1.13688  | 0.77826 | 1.461  | 0.150584 |     |
| BlockB8:TreatLff | -0.31762 | 0.77826 | -0.408 | 0.685000 |     |
| BlockB2:Treatf   | -0.25544 | 0.77826 | -0.328 | 0.744170 |     |
| BlockB3:Treatf   | -0.78035 | 0.77826 | -1.003 | 0.321038 |     |
| BlockB4:Treatf   | -0.89989 | 0.77826 | -1.156 | 0.253284 |     |
| BlockB5:Treatf   | -0.30425 | 0.77826 | -0.391 | 0.697569 |     |
| BlockB6:Treatf   | -0.78442 | 0.77826 | -1.008 | 0.318549 |     |
| BlockB7:Treatf   | 0.81413  | 0.77826 | 1.046  | 0.300755 |     |
| BlockB8:Treatf   | -0.86358 | 0.77826 | -1.110 | 0.272688 |     |
| BlockB2:Treatff  | -0.48627 | 0.77826 | -0.625 | 0.535050 |     |

```

BlockB3:Treatff -0.69556 0.77826 -0.894 0.375923
BlockB4:Treatff -1.67246 0.77826 -2.149 0.036708 *
BlockB5:Treatff 0.12450 0.77826 0.160 0.873579
BlockB6:Treatff -0.57956 0.77826 -0.745 0.460088
BlockB7:Treatff 0.58699 0.77826 0.754 0.454389
BlockB8:Treatff -0.99454 0.77826 -1.278 0.207428
---
```

Signif. codes:

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.5503 on 48 degrees of freedom  
 Multiple R-squared: 0.9284, Adjusted R-squared: 0.8583  
 F-statistic: 13.24 on 47 and 48 DF, p-value: 7.545e-16

```

> ## Typical ANOVA table, often receives primary interest
> ## in an ANOVA context.
> anova(case1301TC.lm)
```

Analysis of Variance Table

```

Response: log(Cover/(100 - Cover))
          Df Sum Sq Mean Sq F value Pr(>F)
Block      7 76.239 10.8912 35.9634 <2e-16 ***
Treat      5 96.993 19.3986 64.0553 <2e-16 ***
Block:Treat 35 15.230 0.4352  1.4369 0.1209
Residuals  48 14.536 0.3028
---
```

Signif. codes:

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that, aside from the parameter estimates/tests, essentially all of the remaining output, above, is identical to that given for our sum-to-zero effects analysis: again, the cell means model and the saturated, non-additive effects models—either coding—are equivalent models, up to parameter interpretation, i.e., up to reparameterization.

### 10.7.2 E.g.: Effects Trmt ANOVA For Common $C\beta$

We do not repeat the discussion of the ANOVA table that we gave in the context of the sum-to-zero analysis, above. The table is the same (compare `anova` output of the current treatment constraint analysis, above, to that of

the sum-to-zero analysis, given previously, or to [RS13, Display 13.10]). Still, as in the analogous section, above, using sum-to-zero constraints, we warn that ANOVA tables may be ambiguous about what is being tested, in either the **BALANCED** or **UNBALANCED** cases; see Section 10.8, below, for more about this.

### 10.7.3 E.g.: Factor Effects Trmt F v R & $\mathbf{C}\boldsymbol{\beta}$ Approach

As we did for the sum-to-zero case, above, we illustrate the F v R (aka extra sums of squares) and  $\mathbf{C}\boldsymbol{\beta}$  approaches. Of course, results are the same: no interaction. Incidentally, the linear combinations implementation appears to be same as in the sum-to-zero case, above, despite the different parameter interpretations. Why? More discussion in class.

**First: F v R code.** Once again, now with treatment constraints, notice the reduced model ANOVA table given by `anova` shows that the sum-of-squares for the remaining terms, `Block` and `Treat`, are the same as those given by the full model. Also, compare [RS13, Displays 13.10 & 13.11]. Again, that the sum-of-squares, etc., do not change between F and R models generally only happens in **BALANCED** cases. Do you see more clearly now a potential ambiguity of tests as presented in ANOVA tables? More about this in Section 10.8, below.

```
> ## Next is essentially identical to the sum-to-zero analysis above:
> ##
> ## Reduced model, without interaction term:
> case1301TCR.lm<- update(case1301TC.lm,. ~ . - Block:Treat)
> ## Just to illustrate effect of balance, otherwise...
> anova(case1301TCR.lm)
```

Analysis of Variance Table

```
Response: log(Cover/(100 - Cover))
          Df Sum Sq Mean Sq F value    Pr(>F)
Block      7 76.239 10.8912 30.368 < 2.2e-16 ***
Treat      5 96.993 19.3986 54.090 < 2.2e-16 ***
Residuals 83 29.767  0.3586
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> ##...just go to F v R test:
> anova(case1301TCR.lm, case1301TC.lm)

Analysis of Variance Table

Model 1: log(Cover/(100 - Cover)) ~ Block + Treat
Model 2: log(Cover/(100 - Cover)) ~ Block + Treat + Block:Treat
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     83 29.767
2     48 14.536 35      15.23 1.4369 0.1209

> ## which gives the same test for interaction as given
> ## by anova(case1301TC.lm), as promised.
```

Next, we consider the **linear combinations approach**. Which elements of  $\beta$  are associated with the interaction, i.e., with the  $(\alpha\beta)_{ij}$ , i.e., with the **Treat:Block** model formula term? Why is the implementation the same as in the sum-to-zero analysis?

```
> ## I build C matrix in stages (not necessary).
> ##
> a<- 8; b<- 6 ## number of factor levels
> Ca<- matrix(0,nrow=(a-1)*(b-1), ncol=a-1)
> Cb<- matrix(0,nrow=(a-1)*(b-1), ncol=b-1)
> Cab<- diag((a-1)*(b-1))
> Cmat<- cbind(0,Ca,Cb,Cab) ## C matrix
> d<- rep(0,(a-1)*(b-1)) ## null CBeta value
> glh.test(reg=case1301TC.lm, cm=Cmat, d=d)

Test of General Linear Hypothesis
Call:
glh.test(reg = case1301TC.lm, cm = Cmat, d = d)
F = 1.4369, df1 = 35, df2 = 48, p-value = 0.1209

> ## Result is same as F v R approach or anova(case1301TC.lm)
> ## of course!
```

#### 10.7.4 E.g.: Effects Trmt Summary

No matter how we approached the analysis—sum-to-zero or treatment constraints; ANOVA table vs. F v R vs. linear combinations—there is little evidence for serious, significant interaction. Before moving on with inferences about the main effects of grazing in the additive model—see **Strategy for Analysis** above—we discuss some potential ambiguity that may arise when using standard ANOVA tables.

### 10.8 SS Type, Balance & the Marginality Principle

The typical ANOVA table may serve as a convenience to the informed analyst, avoiding an explicit implementation via the F v R or the  $\mathbf{C}\boldsymbol{\beta}$  approaches, but ANOVA tables may also allow for abuse by the novice. This situation stems from the different ***types of sum-of-squares*** that may be reported in the table, from (im)balance, and from marginality (in)considerations.

ANOVA tables generally report automatically one of two types of sums of squares (SS): **type I SS**, aka **sequential SS**, or **type III SS**, aka **partial sums-of-squares**, which, of course, also dictates the tests reported by tables. Thus, it behooves us to know which type of SS a table is using, which type of tests are being reported, if we wish to avoid doing something dumb. What happened to type II (and IV) SS? Nothing. We will not discuss them further except to refer the curious to most any SAS manual on linear models (mostly ANOVA material).

**NOTE:** When using the F v R (“extra SS”) approach or  $\mathbf{C}\boldsymbol{\beta}$  approach, we typically need not concern ourselves with the potentially confusing issues arising from SS type and (im)balance that surround standard ANOVA

tables. This is because these other approaches, while arguably less convenient, require us to be more explicit about what exactly is being tested, which is a good reason why we have been using these other approaches concurrently throughout this course. Still, with any approach, we probably want to consider the marginality principle; see below.

- As we mentioned, above, when discussing the factor effects model with sum-to-zero constraints (§10.6.2), we always have

$$SSTO = SSTR + SSE,$$

i.e.,  $SSTO$  retains an additive decomposition into  $SSTR$  (“among groups” SS or “regression” SS) and  $SSE$ . There is no issue here; on the other hand, there is not much at risk, either: the overall F-test, which is often not of primary interest. (R typically does not report  $SSTO$ , but some ANOVA tables may show it.)

- However, problems may arise when we move beyond the overall F-test to consider the slightly more detailed, but standard inferences about main and interaction effects. These sorts of standard inferences are what the ANOVA table is designed to conveniently summarize, as we’ve seen, above. To reiterate, the prototypical ANOVA table seeks to somehow “decompose” or “Analyze”  $SSTR$  into parts associated with each set of main and interaction effects (i.e., with factor covariates and with their interactions) in a model, so we can somehow assess the significance of each main/interaction effect (factor covariate or their interaction). The F v R and C $\beta$  approaches somehow make these tests more explicit, without the ANOVA table pitfalls, at the price of having to know model details, as illustrated in §10.6.3 and 10.7.3.

### 10.8.1 Sequential SS ANOVA

For **type I SS** (sequential) ANOVA tables, SS values are those that result, for their respective terms, after all *previous* (but not subsequent) terms in the table are in the model. Thus, if `FactorA` is reported first, its SS value may be different than when it is reported on the second line of the ANOVA table after, say, e.g., the `FactorB` line. That is, with suggestive (“extra SS”) notation, we might have an ANOVA table reporting SS as

$$\begin{aligned} & SS(A) \\ & SS(B|A) \\ & SS(AB|A, B) \end{aligned}$$

or as

$$\begin{aligned} & SS(B) \\ & SS(A|B) \\ & SS(BA|B, A) \end{aligned}$$

so that

$$\begin{aligned} SSTR &= SS(A) + SS(B|A) + SS(AB|A, B) \\ &= SS(B) + SS(A|B) + SS(BA|B, A), \end{aligned}$$

but, generally,

$$\begin{aligned} SS(A) &\neq SS(A|B) \quad \text{and} \\ SS(B) &\neq SS(B|A) \end{aligned}$$

The last SS components are the same if they are for the same term. For example, the interaction term enters last in both cases, above, so that  $SS(BA|B, A) = SS(AB|A, B)$ .

In the **balanced** case, there is no such dependence of SS values on the term order in an ANOVA table (sequential type I or otherwise), so we didn’t have to talk about this dependence on order of appearance in the table (type I or otherwise) in our running example of the balanced intertidal seaweed grazing experiment (§10.6.2).

To illustrate this ambiguity, I create an **unbalanced version of the data set** from our running example for seaweed grazers.

```
> ## Balanced: nij = n = 2
> with(case1301.df, table(interaction(Block, Treat)))
```

|        |        |        |        |        |        |        |        |
|--------|--------|--------|--------|--------|--------|--------|--------|
| B1.C   | B2.C   | B3.C   | B4.C   | B5.C   | B6.C   | B7.C   | B8.C   |
| 2      | 2      | 2      | 2      | 2      | 2      | 2      | 2      |
| B1.L   | B2.L   | B3.L   | B4.L   | B5.L   | B6.L   | B7.L   | B8.L   |
| 2      | 2      | 2      | 2      | 2      | 2      | 2      | 2      |
| B1.Lf  | B2.Lf  | B3.Lf  | B4.Lf  | B5.Lf  | B6.Lf  | B7.Lf  | B8.Lf  |
| 2      | 2      | 2      | 2      | 2      | 2      | 2      | 2      |
| B1.LfF | B2.LfF | B3.LfF | B4.LfF | B5.LfF | B6.LfF | B7.LfF | B8.LfF |
| 2      | 2      | 2      | 2      | 2      | 2      | 2      | 2      |
| B1.f   | B2.f   | B3.f   | B4.f   | B5.f   | B6.f   | B7.f   | B8.f   |
| 2      | 2      | 2      | 2      | 2      | 2      | 2      | 2      |
| B1.ffF | B2.ffF | B3.ffF | B4.ffF | B5.ffF | B6.ffF | B7.ffF | B8.ffF |
| 2      | 2      | 2      | 2      | 2      | 2      | 2      | 2      |

```
> ## Randomly omit 1 observation from each of half of the treatment
> ## levels to create imbalance (a*b = 8*6 = 48 treatments: randomly
> ## choose 24 treatments from which to omit 1 of 2 observations)
> set.seed(24601)
> indx<- sample(seq(1,95,2), size=24) + sample(0:1,size=24, repl=TRUE)
> dim(unbal.df<- case1301.df[-indx,])
[1] 72 3
> with(unbal.df, table(interaction(Block, Treat)))
```

|        |        |        |        |        |        |        |        |
|--------|--------|--------|--------|--------|--------|--------|--------|
| B1.C   | B2.C   | B3.C   | B4.C   | B5.C   | B6.C   | B7.C   | B8.C   |
| 1      | 1      | 2      | 1      | 1      | 2      | 1      | 1      |
| B1.L   | B2.L   | B3.L   | B4.L   | B5.L   | B6.L   | B7.L   | B8.L   |
| 2      | 1      | 1      | 2      | 1      | 1      | 2      | 1      |
| B1.Lf  | B2.Lf  | B3.Lf  | B4.Lf  | B5.Lf  | B6.Lf  | B7.Lf  | B8.Lf  |
| 2      | 1      | 2      | 2      | 2      | 2      | 1      | 1      |
| B1.LfF | B2.LfF | B3.LfF | B4.LfF | B5.LfF | B6.LfF | B7.LfF | B8.LfF |
| 1      | 2      | 2      | 2      | 1      | 1      | 2      | 2      |
| B1.f   | B2.f   | B3.f   | B4.f   | B5.f   | B6.f   | B7.f   | B8.f   |
| 2      | 2      | 2      | 1      | 1      | 1      | 2      | 2      |
| B1.ffF | B2.ffF | B3.ffF | B4.ffF | B5.ffF | B6.ffF | B7.ffF | B8.ffF |
| 1      | 1      | 2      | 2      | 1      | 2      | 1      | 2      |

Now, for example, I can specify the same model in R as

$$R \sim A + B + A:B,$$

implying  $SSTR = SS(A) + SS(B|A) + SS(AB|A, B)$ , or as

$$R \sim B + A + B:A,$$

implying  $SSTR = SS(B) + SS(A|B) + SS(BA|B, A)$ . The differences between  $SS(A)$  and  $SS(A|B)$  or between  $SS(B)$  and  $SS(B|A)$  are not dramatic in this particular case, but you can see the change; and, as mentioned,  $SS(AB|A, B) = SS(BA|B, A)$ , as these enter the type I sequential table last in each case. (The different specifications result in different permutations of the  $\beta$  vector elements, but this is not relevant in the current context as we are not explicitly considering “picking-off” parameters with  $C$  matrices; again, ANOVA tables can be convenient...)

```
> ##
> ## Unbalanced: Block ('`Factor A'') first, then Treat ('`B''):
> ##
> anova(unbal1.lm<- lm(log(Cover/(100-Cover)) ~
+                         Block + Treat + Block:Treat,
+                         data=unbal.df))
```

Analysis of Variance Table

```
Response: log(Cover/(100 - Cover))
          Df Sum Sq Mean Sq F value    Pr(>F)
Block       7 45.487  6.4981 18.4155 3.180e-08 ***
Treat       5 78.921 15.7842 44.7321 2.161e-11 ***
Block:Treat 35 11.705  0.3344  0.9478     0.5658
Residuals   24  8.469  0.3529
---
```

Signif. codes:

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> ##
> ## Unbalanced: Treat ('`Factor B'') first, then Block ('`A''):
> ##
> anova(unbal2.lm<- lm(log(Cover/(100-Cover)) ~
+                         Treat + Block + Treat:Block,
+                         data=unbal.df))
```

## Analysis of Variance Table

```
Response: log(Cover/(100 - Cover))
          Df Sum Sq Mean Sq F value    Pr(>F)
Treat      5  78.901 15.7802 44.7209 2.167e-11 ***
Block      7  45.506  6.5009 18.4235 3.167e-08 ***
Treat:Block 35 11.705  0.3344  0.9478     0.5658
Residuals  24  8.469  0.3529
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, the question is, ***which test(s) do we use?***

### 10.8.2 Partial SS ANOVA

Perhaps in an attempt to remove any ambiguity about which decomposition to use for tests, **type III SS (partial)** are the SS that result for a (main effect or interaction) term **after all other** terms are in the model. Generally speaking, these are the types of tests that we want to do (and that we have been doing in regression and ANOVA all along): test for the significance of a term(s) after all others are in the model: partial F tests (or partial t tests). **BUT**, likely, we do not want to violate the **marginality** principle.

So, no matter what order the terms/SS occur in a type III (partial) SS ANOVA table, the SS and tests are (numerically) the same for each term; we might write suggestively, e.g.,

$$\begin{aligned} & SS(A|B, AB) \\ & SS(B|A, AB) \\ & SS(AB|A, B) \end{aligned}$$

We use the **Anova** function (capital A) the the **car** library to illustrate. Of course, as mentioned, the type III ANOVAs are the same, regardless of the order of terms, e.g.,

```
> library(car)
> Anova(unbal1.lm, type="III")

Anova Table (Type III tests)
```

```

Response: log(Cover/(100 - Cover))
           Sum Sq Df F value    Pr(>F)
(Intercept) 1.4600  1  4.1377 0.053134 .
Block        10.7421  7  4.3490 0.003085 **
Treat        7.4056  5  4.1974 0.006984 **
Block:Treat 11.7049 35  0.9478 0.565756
Residuals   8.4686 24
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> Anova(unbal2.lm, type="III")

Anova Table (Type III tests)

Response: log(Cover/(100 - Cover))
           Sum Sq Df F value    Pr(>F)
(Intercept) 1.4600  1  4.1377 0.053134 .
Treat        7.4056  5  4.1974 0.006984 **
Block        10.7421  7  4.3490 0.003085 **
Treat:Block 11.7049 35  0.9478 0.565756
Residuals   8.4686 24
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Again,  $SS(AB|A, B)$  (or  $SS(BA|B, A)$ ) is the same in either the type I or type III tables because the interaction term enters last.

But now, the type III table somehow encourages us (well, perhaps a novice) to test a main effect before testing the higher order interaction that contains the main effect term. A bit more in class.

### 10.8.3 Marginality Principle

We had the same issue with the typical regression summary (from the `summary` function) that reports partial regression coefficients and partial t-tests, and, then, too, we warned that we almost always want to obey the ***marginality principle: briefly, don't test lower order terms that exist in higher order terms that are still in the model***. The fact that R's "default"

`anova` function presents only type I (sequential) SS *may* be seen as an (deliberate?) attempt to discourage violation of the marginality principle: we would somehow have to get `anova` to put the desired term to be tested as the last entry in the table (which is the only entry that presents a type III (partial) F/t test in a sequential table as given by `anova`.)

So, in an apparently deliberate attempt to violate the marginality principle by testing main effects before interaction effects (and, being unaware of `cars` `Anova` function), we might try something like

$$R \sim A:B + A + B ,$$

or

$$R \sim B:A + B + A.$$

Notice, however, in the output, below, while `anova` does change the table entry order of the similar (main effects) terms, A and B, it does not allow the higher order term to come before the lower order terms! Thus, `anova` (not `Anova`) seems to discourage attempts to test lower order terms before higher order terms that contain them.

```
> ##
> ##
> anova(unbal1.lm<- lm(log(Cover/(100-Cover)) ~
+                         Block:Treat + Block + Treat,
+                         data=unbal.df))

Analysis of Variance Table

Response: log(Cover/(100 - Cover))
          Df  Sum Sq Mean Sq F value    Pr(>F)
Block       7  45.487  6.4981 18.4155 3.180e-08 ***
Treat       5  78.921 15.7842 44.7321 2.161e-11 ***
Block:Treat 35 11.705  0.3344  0.9478     0.5658
Residuals   24  8.469  0.3529
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> ##
> ##
```

```
> anova(unbal2.lm<- lm(log(Cover/(100-Cover)) ~
+                               Treat:Block + Treat + Block,
+                               data=unbal.df))
```

Analysis of Variance Table

```
Response: log(Cover/(100 - Cover))
          Df Sum Sq Mean Sq F value    Pr(>F)
Treat      5 78.901 15.7802 44.7209 2.167e-11 ***
Block      7 45.506  6.5009 18.4235 3.167e-08 ***
Treat:Block 35 11.705  0.3344  0.9478     0.5658
Residuals  24  8.469  0.3529
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R also has the `drop1` function, which gives partial F tests (when asked) without, evidently, violating the marginality principle. Notice, in the output, below, we get tests that result by “dropping” each term, individually, while leaving all other terms in the model, only if the test does not violate marginality; there are no main effects tests reported because there is an interaction term in the model:

```
> drop1(unbal1.lm, test="F") ## no Block or Treat test: good for marginality
Single term deletions

Model:
log(Cover/(100 - Cover)) ~ Block:Treat + Block + Treat
                           Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>                      8.4686 -58.101
Block:Treat 35       11.705 20.1735 -65.605  0.9478 0.5658

> drop1(unbal2.lm, test="F") ## no Block or Treat test: good for marginality
Single term deletions

Model:
log(Cover/(100 - Cover)) ~ Treat:Block + Treat + Block
                           Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>                      8.4686 -58.101
Treat:Block 35       11.705 20.1735 -65.605  0.9478 0.5658
```

You can force `drop1` to ignore (higher order) terms by using the `scope` option and risk violating the marginality principle. Notice the tests for main effects below are the same as given in the type III (partial) ANOVAs, above, given by `car`'s `Anova` function. To be sure, we are violating the marginality principle here by testing main effects with a relevant higher order interaction in the model.

```
> drop1(unbal1.lm, scope= ~ Block + Treat, test="F")

Single term deletions

Model:
log(Cover/(100 - Cover)) ~ Block:Treat + Block + Treat
      Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>          8.4686 -58.101
Block    7   10.7421 19.2107 -13.126  4.3490 0.003085 ***
Treat    5    7.4056 15.8742 -22.862  4.1974 0.006984 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> drop1(unbal2.lm, scope= ~ Treat + Block, test="F")

Single term deletions

Model:
log(Cover/(100 - Cover)) ~ Treat:Block + Treat + Block
      Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>          8.4686 -58.101
Treat    5    7.4056 15.8742 -22.862  4.1974 0.006984 ***
Block    7   10.7421 19.2107 -13.126  4.3490 0.003085 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we have illustrated some time ago, when discussing interaction plots and components of ANOVA (Section 10.2), in addition to the current discussion, the interpretation of main effects in the presence of interaction is often non-sensical (e.g., what good is it to say that A (B) has no (main) effect when, in fact, A (B) may have very interesting effects depending on which level of B (A) we are in). Thus, as we have proceeded (again, see bullet

“Strategy for Analysis,” above), we first investigate our model to see if we can reasonably omit the interaction term in an attempt not to endanger our interpretation of main effects.

This practice is consistent with the principle of marginality, which, again, “requires” that we retain all lower order terms in the presence of higher order terms: e.g., if AB interaction is in the model, then A and B factors should be in the model. That is, we proceeded by first investigating the higher order terms before, possibly, throwing them out and moving to a model with lower order terms—an additive model in this two-way discussion. Moreover, in keeping with the marginality principle, the “Strategy” offered does not tell us to attempt removal of lower order terms if we cannot reasonably remove higher order terms before moving on to more detailed analysis.

As mentioned, above, the marginality principle is applicable in regression, too: if, e.g.,  $xy$  is in the model then  $x$  and  $y$  (and 1) should be in the model; similarly for  $x^2$  and  $x$  (and 1). See [KNNL05, pg. 299] for their “Hierarchical Approach to Fitting,” which is essentially a description of the marginality principle in practice for regression models.

Finally, note again that our F v R or  $C\beta$  approaches do not invite much confusion regarding what is being tested, but this comes at the expense of having to think more explicitly about what your tests are and how to implement the tests in R.

#### 10.8.4 Balance

Before leaving this section, we note how **balance** may contribute to confusion about which test is being performed. As mentioned, when our treatment design is balanced, there is no dependency of the SS on the order in which terms occur in an ANOVA table (sequential (type I) or partial (type III)). In other words, the sequential (type I) and partial (type III) ANOVA tables are the same, no matter which order the terms are presented in the tables.

But, then, we face the question of ***which test is being performed?***

Is the line for, e.g., **Treat** testing for the effect of treatment after **Block** and **Block:Treat** are in the model? Or, is it testing for the effect of **Treat** after only **Block** is in the model? Or, after only the intercept? Which is

it? After all, the value of the SS for **Treat** (or other terms) never changes regardless of its order in the ANOVA table or regardless of any other terms that may or may not be in the model.

Typically, we test for a term's effects after all others are in the model, which is what we have been doing all along for regression and ANOVA. So, despite the SS for a term always being the same regardless of table order or regardless of other terms in/out of the model (in the balanced case!), we typically want to think that we're doing a partial F test for a term after all other terms are in the model. And, again, of course, we likely want to obey the marginality principle, too.

## 10.9 Additive Model: Tests for Overall Main Effects

After investigating interaction and deciding non-significance, a typical course of action is to proceed with the additive model to make further inference about main effects, e.g., grazing effects or block effects in our running example. See the bullet, “**Strategy for Analysis**,” above. Again, see §10.2 for how the presence of interactions can mislead interpretation of main effects.

Next, we look to see if we have significant effects of locations (blocks). Yes, clearly we do, as indicated by the test for the **Block** term in any (transformed) ANOVA table given so far (not using the unbalanced data discussed above!). Why can we use any table given so far?

In keeping with our pattern of presentation, in addition to the ANOVA tables, we can, of course, use the F v R or linear combinations approach. Now, however, our “F”ull model is now the additive model, without interaction effects.

### 10.9.1 F v R Approach

First, we present the F v R approach to testing for significant block effects.

```
> #####
> ## Sum-to-zero analysis:
> #####
> ## New reduced model:
> case1301aR2.lm<- update(case1301aR.lm, . ~ . - Block)
```

```
> ## Just to illustrate balance (non) effect on Treat SS...:
> anova(case1301aR2.lm)
```

Analysis of Variance Table

```
Response: log(Cover/(100 - Cover))
          Df  Sum Sq Mean Sq F value    Pr(>F)
Treat      5   96.993 19.3986   16.47 1.625e-11 ***
Residuals 90  106.005  1.1778
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> ##...otherwise go directly to F v R test:
> anova(case1301aR2.lm, case1301aR.lm)
```

Analysis of Variance Table

```
Model 1: log(Cover/(100 - Cover)) ~ Treat
Model 2: log(Cover/(100 - Cover)) ~ Block + Treat
          Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1       90 106.005
2       83 29.767  7    76.239 30.368 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> ## or, why not this (note use of fullest model MSE)?:
> anova(case1301aR2.lm, case1301aR.lm, case1301a.lm)
```

Analysis of Variance Table

```
Model 1: log(Cover/(100 - Cover)) ~ Treat
Model 2: log(Cover/(100 - Cover)) ~ Block + Treat
Model 3: log(Cover/(100 - Cover)) ~ Block + Treat + Block:Treat
          Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1       90 106.005
2       83 29.767  7    76.239 35.9634 <2e-16 ***
3       48 14.536 35    15.230  1.4369 0.1209
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> #####
> ## Treatment analysis:
> #####
> ## New reduced model:
> case1301TCR2.lm<- update(case1301TCR.lm, . ~ . - Block)
> ## Just to illustrate balance (non) effect on Treat SS...:
> anova(case1301TCR2.lm)
```

Analysis of Variance Table

```
Response: log(Cover/(100 - Cover))
          Df  Sum Sq Mean Sq F value    Pr(>F)
Treat      5   96.993 19.3986   16.47 1.625e-11 ***
Residuals 90  106.005  1.1778
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> ##...otherwise go directly to F v R test:
> anova(case1301TCR2.lm, case1301TCR.lm)
```

Analysis of Variance Table

```
Model 1: log(Cover/(100 - Cover)) ~ Treat
Model 2: log(Cover/(100 - Cover)) ~ Block + Treat
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     90 106.005
2     83  29.767  7   76.239 30.368 < 2.2e-16 ***
---
Signif. codes:
```

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> ## or, why not this (note use of fullest model MSE)?:
> anova(case1301TCR2.lm, case1301TCR.lm, case1301TC.lm)
```

Analysis of Variance Table

```
Model 1: log(Cover/(100 - Cover)) ~ Treat
Model 2: log(Cover/(100 - Cover)) ~ Block + Treat
Model 3: log(Cover/(100 - Cover)) ~ Block + Treat + Block:Treat
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     90 106.005
2     83  29.767  7   76.239 35.9634 <2e-16 ***
```

```

3      48  14.536 35     15.230  1.4369 0.1209
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

As we already knew, blocks appear to account for significant response variability, something that the researchers suspected and were, presumably, not be surprised by. We keep blocks in our additive model. See [RS13, pg 397].

### 10.9.2 $C\beta$ Approach

Next, we consider the linear combinations approach. Despite the different parameter interpretation between the sum-to-zero analysis and the treatment analysis, the implementation appears to be the same in either case. (Recall that implementations for testing interaction effects also appeared to be the same between these analyses.) Why?

```

> ## Linear combinations approach:
> ##
> a<- 8; b<- 6 ## number of factor levels
> Ca<- diag(a-1)
> Cb<- matrix(0, nrow=(a-1), ncol=b-1)
> Cmat<- cbind(0,Ca,Cb) ## C matrix
> d<- rep(0,(a-1)) ## null CBeta value
> ##
> ## Sum-to-zero analysis:
> ##
> glh.test(reg=case1301aR.lm, cm=Cmat, d=d)

```

```

Test of General Linear Hypothesis
Call:
glh.test(reg = case1301aR.lm, cm = Cmat, d = d)
F = 30.3684, df1 = 7, df2 = 83, p-value = < 2.2e-16

> ##
> ## Treatment analysis:
> ##
> glh.test(reg=case1301TCR.lm, cm=Cmat, d=d)

```

```
Test of General Linear Hypothesis
Call:
glh.test(reg = case1301TCR.lm, cm = Cmat, d = d)
F = 30.3684, df1 = 7, df2 = 83, p-value = < 2.2e-16
```

Now, we move on to specific questions about grazing effects using the additive model with `Block` and `Treat` terms.

## 10.10 Additive Model: More Detailed Inference of Main Effects

Again, as with the tests of main effects (`Block` or `Treat`), in the previous section, even more detailed inferences about main effects generally only make sense in the absence of (bad) interaction.

We will see that, without interaction, the (additive) effects model (either set of constraints) make for relatively straightforward investigation of detailed questions about main effects: such questions typically involve, wonder of wonders, only the main effects parameters of the effect in question: e.g., questions about the effects of factor A (B) typically involve only the  $\alpha_i$  ( $\beta_j$ ), with some attention required for the particular constraints being used; we might also expect an overall effect parameter,  $\mu..$ , to play a role in such questions, but we should not expect a mix of  $\alpha$ s and  $\beta$ s when asking about one factor or another (not both) in a additive model.

Clearly, there are significant, overall block and grazing effects. Again, we mentioned that we do not typically have interest in making further, more detailed inference about block effects, but we typically would want to investigate more detailed inferences about grazing effects, the “treatment” of interest. We follow [RS13, Sections 13.3.4 & 13.3.5] with the R code to implement (some of) the first two of their inferences, numbered 1-5 in [RS13, Section 13.3.4]. We will have some discussion in class. You may be asked about the remaining inferences on a homework. See [RS13, Sections 13.3.4 & 13.3.5] for more.

To help us keep track of factor levels, let’s refer to th “Grazers Allowed” table in [RS13, Display 13.7], which we reproduced, above. This table is

merely helping us to interpret the levels of our factor B, the levels of **Treat**, the grazing “treatment” (technically, factor) to enable further analysis among these levels. In particular, we will re-order our factor B (**Treat**) levels to match the order shown in that table; see code below.

Note that R&S make little mention of factor A (**Block**) at this point in their discussion. We’ve decided that the effects of **Treat** do not depend on (do not interact with) **Block** and we have kept the significant but otherwise relatively uninteresting **Block** (location) effects in the model to help elucidate **Treat** effects or some more detailed and interesting linear combinations thereof; again, the additive model is convenient for subsequent, more detailed inference for the **Treat** factor. (Apologies for constantly repeating such things!)

- **Do large fish have an effect on the regeneration ratio?** If so, how much?

The table suggests that the following linear combination (specifically, contrast) of (marginal factor) means embodies our question:

$$\gamma_1 = \frac{1}{2}(\mu_{fF} - \mu_f) + \frac{1}{2}(\mu_{LfF} - \mu_{Lf}),$$

which is the average effect of large fish (F), averaged over the case when small fish are present (f), without limpets(L), and the case when small fish (f) and limpets (L) are present. See the summary of linear combination coefficients in [RS13, Display 13.13].

Given that we will order the levels of the **Treat** factor to match their order in [RS13, Display 13.7], we can write, in our notation,

$$\gamma_1 = \frac{1}{2}(\mu_{.3} - \mu_{.2}) + \frac{1}{2}(\mu_{.6} - \mu_{.5}),$$

which is more explicit about the fact that we are averaging over blocks.

While using means to specify detailed linear combinations may be convenient, especially when interactions exist (not here), we translate  $\gamma_1$  to our effects model(s). In either the sum-to-zero or treatment constraints cases, we have

$$\beta_j = \mu_{.j} - \mu_{..},$$

This allows us to write the above linear combination in terms of the treatment effects,  $\beta_j$ :

$$\gamma_1 = \frac{1}{2}(\beta_3 - \beta_2) + \frac{1}{2}(\beta_6 - \beta_5).$$

Or, perhaps we should write,

$$\gamma_1 = \frac{1}{2}(\beta_{fF} - \beta_f) + \frac{1}{2}(\beta_{LfF} - \beta_{Lf}).$$

As mentioned above, questions about main effects, in the absence of interaction, typically resolve to sensible linear combinations of main effects parameters, but, more generally, we may expect to see an overall effect  $\mu_{..}$ . Because we have a contrast, here, the overall effect goes away.

To proceed further, we must consider which constraint(s) we are using, because the interpretation of the  $\beta_j$  and the (non-redundant)  $\boldsymbol{\beta}$  vector depend on the constraints used, as we know. First, let's consider the treatment constraints/coding, which is the default in R. It turns out that this is the more convenient coding, in some sense, for question 1. Why? More in class.

For the treatment constraint, recall  $\beta_1 = 0$  so that  $\boldsymbol{\beta}$  contains (among others) all  $\beta_j$ ,  $\beta_2$  to  $\beta_b$ , but not  $\beta_1$ . Notice that this requires no further changes to  $\gamma_1$ , i.e., we have  $\gamma_1$  in terms of the elements of the treatment constraints'  $\boldsymbol{\beta}$  vector:

$$\gamma_1 = \frac{1}{2}(\beta_3 - \beta_2) + \frac{1}{2}(\beta_6 - \beta_5) \quad \text{treatment.}$$

(Again, refer to the table in [RS13, Display 13.7] for how we're thinking of the order of factor levels.)

Recall, however, the sum-to-zero constraints'  $\boldsymbol{\beta}$  contains  $\beta_1$  to  $\beta_{b-1}$ , because, as specified,  $\beta_b = -\sum_{j=1}^{b-1} \beta_j$ . Plugging in this definition of  $\beta_6$  ( $b = 6$ ) in the above expression for  $\gamma_1$  gives

$$-\frac{1}{2}\beta_1 - \beta_2 - \frac{1}{2}\beta_4 - \beta_5 \quad \text{sum-to-zero.}$$

```

> ## Are we sure we know how R orders factor levels? help(factor) WARNS:
> ## The levels of a factor are by default sorted, but the sort order
> ## may well depend on the locale at the time of creation, and should
> ## not be assumed to be ASCII.
> levels(case1301.df$Treat)

[1] "C"     "L"     "Lf"    "Lff"   "f"     "ff"

> ## Let's reorder/refit to match Display 13.7 in R&S and their
> ## discussion to help avoid confusion:
> case1301.df$Treat<- factor(case1301.df$Treat,
+                                 levels=levels(case1301.df$Treat)[c(1,5,6,2,3,4)])
> levels(case1301.df$Treat) ## now matches Display 13.7

[1] "C"     "f"     "ff"    "L"     "Lf"    "Lff"

> ## Are we sure which constraint(s) is/are in effect?:
> sapply(case1301.df, attr, which="contrasts")

$Cover
NULL

$Block
  B2 B3 B4 B5 B6 B7 B8
B1  0  0  0  0  0  0  0
B2  1  0  0  0  0  0  0
B3  0  1  0  0  0  0  0
B4  0  0  1  0  0  0  0
B5  0  0  0  1  0  0  0
B6  0  0  0  0  1  0  0
B7  0  0  0  0  0  1  0
B8  0  0  0  0  0  0  1

$Treat
NULL

>getOption("contrasts") ## ...okay to proceed?

[1] "contr.treatment" "contr.treatment"

> ##
> #####
> ## Treatment analysis:

```

```

> #####
> ## Overwrites previous fit of same name (levels okay now?):
> case1301TCR.lm<- lm(log(Cover/(100-Cover)) ~ Block + Treat,
+                         data=case1301.df)
> a<- 8; b<- 6 ## number of factor levels
> Ca<- rep(0,a-1) ## blocks
> Cb<- 1/2 * c(-1, 1, 0, -1, 1) ## grazing
> Cmat<- c(0,Ca,Cb) ## C matrix (vector)
> d<- 0 ## null CBeta value
> glh.test(reg=case1301TCR.lm, cm=Cmat, d=d)

Test of General Linear Hypothesis
Call:
glh.test(reg = case1301TCR.lm, cm = Cmat, d = d)
F = 16.8205, df1 = 1, df2 = 83, p-value = 9.54e-05

> ## Or, somewhat ugly, but readable and perhaps instructive
> ## (compare to ``Contrast Summary'' in Display 13.13 in R&S):
> ##
> estimable(obj=case1301TCR.lm, cm=Cmat, beta0=d, conf.int=0.95)

              beta0   Estimate
(0 0 0 0 0 0 0 -0.5 0.5 0 -0.5 0.5)      0 -0.6140257
                                         Std. Error   t value
(0 0 0 0 0 0 0 -0.5 0.5 0 -0.5 0.5)  0.1497157 -4.101278
                                         DF   Pr(>|t|)
(0 0 0 0 0 0 0 -0.5 0.5 0 -0.5 0.5) 83 9.539869e-05
                                         Lower.CI   Upper.CI
(0 0 0 0 0 0 0 -0.5 0.5 0 -0.5 0.5) -0.9118042 -0.3162472

> ## (Alas, R&S reorder the ``Treatment'' (factor B) levels in
> ## their Display 13.13 of results! Our results match theirs
> ## of course.)
> ##
> #####
> ## Sum-to-zero analysis:
> #####
> contrasts(case1301.df$Block)<- contr.sum(levels(case1301.df$Block))
> contrasts(case1301.df$Treat)<- contr.sum(levels(case1301.df$Treat))
> ## Overwrites previous fit of same name (levels okay now?):
> case1301aR.lm<- lm(log(Cover/(100-Cover)) ~ Block + Treat,
+                         data=case1301.df)

```

```

> Cb<- -c(1/2, 1, 0, 1/2, 1) ## grazing
> Cmat<- c(0,Ca,Cb) ## C matrix (vector)
> glh.test(reg=case1301aR.lm, cm=Cmat, d=d)

Test of General Linear Hypothesis
Call:
glh.test(reg = case1301aR.lm, cm = Cmat, d = d)
F = 16.8205, df1 = 1, df2 = 83, p-value = 9.54e-05

> estimable(obj=case1301aR.lm, cm=Cmat, beta0=d, conf.int=0.95)

              beta0   Estimate
(0 0 0 0 0 0 0 -0.5 -1 0 -0.5 -1)      0 -0.6140257
                                         Std. Error   t value DF
(0 0 0 0 0 0 0 -0.5 -1 0 -0.5 -1)  0.1497157 -4.101278 83
                                         Pr(>|t|)   Lower.CI
(0 0 0 0 0 0 0 -0.5 -1 0 -0.5 -1) 9.539869e-05 -0.9118042
                                         Upper.CI
(0 0 0 0 0 0 0 -0.5 -1 0 -0.5 -1) -0.3162472

> ## Results are the same for either coding, of course.

```

Questions 2 and 3 [RS13, Sec. 13.3.4, Display 13.13] are similar to 1. (Homework?) Let's implement the 4th question to continue our running example. Question 5 is similar to 4. (Homework?) Again, we follow the factor level order given in the table of [RS13, Display 13.7].

- **Do limpets have a different effect when small fish are present than when small fish are absent?**

If we compare the question to the table in [RS13, Display 13.7], we can see two effects of limpets in the presence of small fish (following R&S subscripts for the moment):

$$\mu_{LfF} - \mu_{fF}$$

and

$$\mu_{Lf} - \mu_{f.}$$

We might average these two effects (i.e., average over large fish(L)) to get an average effect of limpets when small fish are present.

We also can see the (single) effect of limpets in the absence of small fish:

$$\mu_L - \mu_C.$$

Thus, we consider the difference of the above (average and single) effects to suffice to answer our question:

$$\gamma_4 = \frac{1}{2}(\mu_{LfF} - \mu_{fF}) + \frac{1}{2}(\mu_{Lf} - \mu_f) - (\mu_L - \mu_C).$$

You may be able to see how this translates to  $\beta_j$  effects notation already. But, we go through maringal means first, as with question 1, to be careful.

$$\gamma_4 = \frac{1}{2}(\mu_{.6} - \mu_{.3}) + \frac{1}{2}(\mu_{.5} - \mu_{.2}) - (\mu_{.4} - \mu_{.1}).$$

As we did for  $\gamma_1$ , we plug in the definition of the  $\beta_j$  effect,

$$\beta_j = \mu_{.j} - \mu_{..},$$

to get

$$\gamma_4 = \frac{1}{2}(\beta_6 - \beta_3) + \frac{1}{2}(\beta_5 - \beta_2) - (\beta_4 - \beta_1).$$

Once again, our contrast annihilates the overall effect,  $\mu_{..}$ , and we should see how we may have gone directly to this last expression for  $\gamma_4$ , in terms of  $\beta_j$ , bypassing consideration of marginal means.

Perhaps we should write,

$$\gamma_4 = \frac{1}{2}(\beta_{LfF} - \beta_{fF}) + \frac{1}{2}(\beta_{Lf} - \beta_f) - (\beta_L - \beta_C).$$

Now, as in question 1, to proceed further, we must consider the particular constraints/coding of the effects model.

For the treatment constraint, recall  $\beta_1 = 0$  so that  $\beta$  contains (among others) all  $\beta_j$ ,  $\beta_2$  to  $\beta_b$ , but not  $\beta_1$ . Unlike the treatment constraint case for question 1, we have some work to do, but not much:

$$\gamma_4 = \frac{1}{2}(\beta_6 - \beta_3) + \frac{1}{2}(\beta_5 - \beta_2) - \beta_4 \quad \text{treatment.}$$

The sum-to-zero constraints'  $\beta$  contains  $\beta_1$  to  $\beta_{b-1}$ , because, as specified,  $\beta_b = -\sum_{j=1}^{b-1} \beta_j$  ( $b = 6$ ). Thus,

$$\gamma_4 = \frac{1}{2}\beta_1 - \beta_2 - \beta_3 - \frac{3}{2}\beta_4 \quad \text{sum-to-zero.}$$

```

> ## Are we sure we know how R orders factor levels?
> ## Careful that the ordering is in the fitted object!
> levels(case1301.df$Treat) ## reflects changes above

[1] "C"    "f"    "fF"   "L"    "Lf"   "LfF"

> #####
> ## Treatment analysis:
> #####
> case1301TCR.lm ## ok, level order looks good

Call:
lm(formula = log(Cover/(100 - Cover)) ~ Block + Treat, data = case1301.df)

Coefficients:
(Intercept)      BlockB2      BlockB3      BlockB4
-1.2226        0.4600       2.1046       2.9807
BlockB5        BlockB6      BlockB7      BlockB8
  1.2160        2.0251       1.1085       1.3300
Treatf        Treatff      TreatL      TreatLf
 -0.4941       -1.0019      -1.8925      -2.1849
TreatLff
 -2.9052

> case1301TCR.lm$contrasts ## 1st time use of this

$Block
  B2 B3 B4 B5 B6 B7 B8
B1  0  0  0  0  0  0  0
B2  1  0  0  0  0  0  0
B3  0  1  0  0  0  0  0
B4  0  0  1  0  0  0  0
B5  0  0  0  1  0  0  0
B6  0  0  0  0  1  0  0
B7  0  0  0  0  0  1  0
B8  0  0  0  0  0  0  1

$Treat
[1] "contr.treatment"

> a<- 8; b<- 6 ## number of factor levels
> Ca<- rep(0,a-1) ## blocks

```

```

> Cb<- 1/2 * c(-1, -1, -2, 1, 1) ## grazing
> Cmat<- c(0,Ca,Cb) ## C matrix (vector)
> d<- 0 ## null CBeta value
> glh.test(reg=case1301TCR.lm, cm=Cmat, d=d)

Test of General Linear Hypothesis
Call:
glh.test(reg = case1301TCR.lm, cm = Cmat, d = d)
F = 0.1356, df1 = 1, df2 = 83, p-value = 0.7136

> ## Or, somewhat ugly, but readable and perhaps instructive
> ## (compare to ``Contrast Summary'' in Display 13.13 in R&S):
> ##
> estimable(obj=case1301TCR.lm, cm=Cmat, beta0=d, conf.int=0.95)

              beta0   Estimate
(0 0 0 0 0 0 0 -0.5 -0.5 -1 0.5 0.5)      0 0.09548527
                                              Std. Error   t value
(0 0 0 0 0 0 0 -0.5 -0.5 -1 0.5 0.5) 0.2593152 0.3682209
                                              DF Pr(>|t|)
(0 0 0 0 0 0 0 -0.5 -0.5 -1 0.5 0.5) 83 0.713646
                                              Lower.CI   Upper.CI
(0 0 0 0 0 0 0 -0.5 -0.5 -1 0.5 0.5) -0.4202822 0.6112528

> ## (Alas, R&S reorder the ``Treatment'' (factor) levels in
> ## Display 13.13!)
> #####
> ## Sum-to-zero analysis:
> #####
> case1301aR.lm$contrasts ## 1st time use of this

$Block
 [,1] [,2] [,3] [,4] [,5] [,6] [,7]
B1     1     0     0     0     0     0     0
B2     0     1     0     0     0     0     0
B3     0     0     1     0     0     0     0
B4     0     0     0     1     0     0     0
B5     0     0     0     0     1     0     0
B6     0     0     0     0     0     1     0
B7     0     0     0     0     0     0     1
B8    -1    -1    -1    -1    -1    -1    -1

```

```
$Treat
 [,1] [,2] [,3] [,4] [,5]
C      1     0     0     0     0
f      0     1     0     0     0
fF     0     0     1     0     0
L      0     0     0     1     0
Lf     0     0     0     0     1
Lff    -1    -1    -1    -1    -1

> ## level order looks good
> ##
> Cb<- c(1/2, -1, -1, -3/2, 0) ## grazing
> Cmat<- c(0,Ca,Cb) ## C matrix (vector)
> glh.test(reg=case1301aR.lm, cm=Cmat, d=d)

Test of General Linear Hypothesis
Call:
glh.test(reg = case1301aR.lm, cm = Cmat, d = d)
F = 0.1356, df1 = 1, df2 = 83, p-value = 0.7136

> estimable(obj=case1301aR.lm, cm=Cmat, beta0=d, conf.int=0.95)

              beta0   Estimate
(0 0 0 0 0 0 0 0.5 -1 -1 -1.5 0) 0 0.09548527
                                         Std. Error t value DF
(0 0 0 0 0 0 0 0.5 -1 -1 -1.5 0) 0.2593152 0.3682209 83
                                         Pr(>|t|) Lower.CI
(0 0 0 0 0 0 0 0.5 -1 -1 -1.5 0) 0.713646 -0.4202822
                                         Upper.CI
(0 0 0 0 0 0 0 0.5 -1 -1 -1.5 0) 0.6112528

> ## Results are the same for either coding, of course
```

Finally, we note that the implementation of the above questions, as linear combinations of parameters, may seem a bit complicated at first. After all, [RS13, Sec. 13.3.4] simply use corresponding averages of observations,  $Y_{ijk}$ , to answer the questions (though they omit much detail)—relatively easy to compute “by hand,” which was relatively important when computational resources were more limited (think: adding machines!). But, their approach of using simple averages generally only works in the **BALANCED** case, as

in the current running example. Their presentation is a hold-over, in my opinion, of the classic balanced presentation of ANOVA that we still see in many textbooks. Our approach, however, works also in **UNBALANCED** cases (assuming there are not other complicating factors, like empty cells, or, more generally, estimability issues). [KNNL05, Chaps. 19-22], too, devote much of their presentation to the balanced case, wherein easy “hand” computations (simple averages) and the classic ANOVA decomposition, which we discussed above, hold, before they move explicitly to unbalanced cases in [KNNL05, Chap 23], often referring to “the regression approach” to ANOVA in such unbalanced cases. [RS13] discuss the regression approach in various sections of their book ([RS13, pg 390, Sec. 13.3.5, 13.4.2, 13.4.3]). I did not check their constraint/coding scheme closely, but it appears to be more like R’s treatment coding than sum-to-zero coding.

In this so-called “regression approach,” the concepts are the same between the balanced case, where special hand computations and additive sum-of-squares ANOVA decomposition apply—as in the seaweed grazing example—and the unbalanced case, where these “features” do not apply in general. (Again, see Section 10.8.) I have made an attempt in these notes to feature the regression approach all along. Essentially, this so-called “regression approach” is the F v R approach (aka extra sum-of-squares approach), where we had to fit a reduced model, explicitly, in addition to full model. As we’ve seen, the fitting of the reduced model is (almost!...ask me to explain in class) not necessary for standard interaction/main effects tests with the standard ANOVA printout in the balanced case. This regression approach may also be called something like “the general linear test” approach, as implemented in `glh.test` or `estimable`, which we’ve tended to call the  $\mathbf{C}\boldsymbol{\beta}$  approach. Whichever implementation to choose—ANOVA table, F v R, or  $\mathbf{C}\boldsymbol{\beta}$ —is mostly a matter of preference.

We can see how the hand computations and additive decomposition of the balanced case were favored when computational resources were relatively limited some time ago. But, nowadays, perhaps with some special exceptions (e.g., large data), it seems to me that we should focus more on the more generally useful “regression approach,” for which, in the vase majority of cases, computational burden is not an issue. Besides, the regression approach

to ANOVA makes explicit the connection to regression—it's just regression on specially coded covariates, and our inferences typically involve some linear combination(s) of the  $\beta$  vector. Moreover, it tends to force us to have a better understanding of software implementation, at least in R.

In any case, I remind you that it is probably good practice to heed the marginality principle, mentioned above, except in rare, special circumstances, which will likely be dictated by (your) science, not by (some gap in your knowledge about) statistics. In other words, don't worry about these special circumstances creating problems by accident.

# Lecture References

- [Ber85] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition, 1985. ISBN 0387960988.
- [Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [CB02] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, Pacific Grove, 2 edition, 2002.
- [GCS<sup>+</sup>14] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, 3rd edition, 2014.
- [Har97] David A. Harville. *Matrix Algebra from a Statistician's Perspective*. Springer–Verlag, 1997.
- [HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
- [HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2 edition, 2009.
- [JWHT14] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.
- [KNNL05] Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models*. McGraw–Hill/Irwin, New York, 5th edition, 2005.

- [Mur12] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
- [Pea09] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.
- [PM18] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, New York, 2018.
- [Rob01] Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, New York, 2001. ISBN 0-387-95231-4.
- [RS13] Fred L. Ramsey and Daniel W. Schafer. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Brooks Cole, Boston, 3rd edition, 2013.
- [RW06] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2006.
- [VR02] W.N. Venables and B.D. Ripley. *Modern applied statistics with S*. Springer-Verlag, New York, 2002.
- [Wak13] Jon Wakefield. *Bayesian and Frequentist Regression Methods*. Springer, New York, 2013.